ARI Research Note 90-126

Assessment of Performance Measurement Methodologies for Collective Military Training

Janet J. Turnage
University of Central Florida

Thomas L. Houser and David A. Hofmann
Institute for Simulation and Training

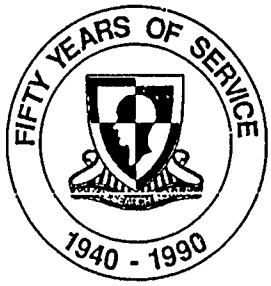
for

Contracting Officer's Representative
Bruce W. Knerr

PM TRADE Field Unit at Orlando, FL
Stephen L. Goldberg, Chief

Training Research Laboratory
Jack H. Hiller, Director

September 1990



United States Army
Research Institute for the Behavioral and Social Sciences

U.S. ARMY RESEARCH INSTITUTE FOR THE BEHAVIORAL AND SOCIAL SCIENCES

A Field Operating Agency Under the Jurisdiction of the Deputy Chief of Staff for Personnel

EDGAR M. JOHNSON
Technical Director

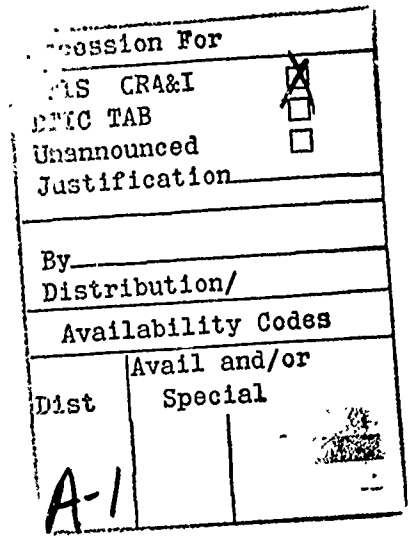
JON W. BLADES
COL, IN
Commanding

Research accomplished under contract for the Department of the Army

University of Central Florida

Technical review by

Larry L. Meliza



NOTICES

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1a. REPORT SECURITY CLASSIFICATION Unclassified	1b. RESTRICTIVE MARKINGS --
2a. SECURITY CLASSIFICATION AUTHORITY --	3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution is unlimited.
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE --	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) --	5. MONITORING ORGANIZATION REPORT NUMBER(S) ARI Research Note 90-126
6a. NAME OF PERFORMING ORGANIZATION University of Central Florida 6b. OFFICE SYMBOL (If applicable) --	7a. NAME OF MONITORING ORGANIZATION U.S. Army Research Institute PM TRADE Field Unit at Orlando, FL
6c. ADDRESS (City, State, and ZIP Code) 12424 Research Parkway, Suite 300 Orlando, FL 32826	7b. ADDRESS (City, State, and ZIP Code) ATTN: PERI-IF 12350 Research Parkway Orlando, FL 32826-3276
8a. NAME OF FUNDING/SPONSORING ORGANIZATION U.S. Army Research Institute for the Behavioral and Social Sciences 8b. OFFICE SYMBOL (If applicable) PERI-I	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER N61339-85-D-0024
8c. ADDRESS (City, State, and ZIP Code) 5001 Eisenhower Avenue Alexandria, VA 22333-5600	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO.: 63007A; PROJECT NO.: 795; TASK NO.: 345; WORK UNIT ACCESSION NO.: C6

11. TITLE (Include Security Classification)

Assessment of Performance Measurement Methodologies for Collective Military Training

12. PERSONAL AUTHOR(S) Turnage, Janet J. (University of Central Florida); Houser, Thomas L.; and Hofmann, David A. (Institute for Simulation and Training)

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Final	FROM 86/04 TO 87/08	1990, September	137

16. SUPPLEMENTARY NOTATION

This report is jointly sponsored by the U.S. Army Research Institute and the DoD Training Performance Data Center, 3280 Progress Drive, Orlando, FL 32826-3229.

17. COSATI CODES		
FIELD	GROUP	SUB-GROUP

18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)

Performance measurement; Training devices; ARTEP; Reliability; Team training; Assessment; Collective training; Training research; (Continued)

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

This report describes the state of the art in military collective performance measurement methodologies, particularly those used in the Army. The research, which is based on a large literature review, covers past, present, and emerging training systems and performance measurement issues. It discusses problems in collective training research, such as inadequate definitions of collective concepts, lack of team development models, and difficulties in differentiating between individual and collective skills. Difficulties specifying objectives, conditions, and standards lead to lack of measurement reliability. Reliability issues are discussed in relation to Army Training and Evaluation Programs (ARTEP) and other current training devices (e.g., SIMNET). The conclusions recommend (1) further study of important dimensions of collective training, (2) utilization of critical incidents methodologies to identify fundamental characteristics of effective collective behaviors, and (3) development of reliable, standardized measurement systems that should test the efficacy of surrogate measurement. Key words: training devices, assessment, simulation, teams

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT	21. ABSTRACT SECURITY CLASSIFICATION
☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS	Unclassified
22a. NAME OF RESPONSIBLE INDIVIDUAL Halim Ozkaptan	22b. TELEPHONE (Include Area Code) (407) 380-8173
	22c. OFFICE SYMBOL PERI-IF

DD Form 1473, JUN 86 Previous editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE

UNCLASSIFIED

ARI Research Note 90-126

18. SUBJECT TERMS (Continued)

Battlefield simulation

ASSESSMENT OF PERFORMANCE MEASUREMENT METHODOLOGIES FOR COLLECTIVE MILITARY TRAINING

EXECUTIVE SUMMARY

Requirement:

To describe the state of the art in collective performance measurement methodologies in order to provide recommendations for future research and development (R&D) aimed at improving performance measurement in Army training. The Army has long recognized that the performance of integrated crews, teams, and units is essential to overall mission success. Despite this, the current state of collective training evaluation has remained at a relatively unsophisticated level. Lack of understanding of the important dimensions of collective training and evaluation has hampered attempts to adequately assess combat readiness. Data are required that will allow qualification of training's impact on readiness.

Procedure:

This instrument is based on a literature review of current measurement systems and methodologies used in both military and nonmilitary settings. It comes in large part from Army Research Institute (ARI) publications. The information collected in this report covers past, present, and emerging training systems and performance measurement issues, with an attempt to focus on the measurement of collective, rather than individual, performance.

Findings:

The report is divided into five chapters that discuss major topical areas and a sixth chapter that summarizes the main research issues and implications. Each chapter can be read independently as an overview of the particular topic covered. Chapter I provides the main documentation of research needs and outlines the approach used to document information contained in the report.

The second chapter reviews the state of the art in team training research. In order for improvement in collective training research to occur, an accepted definition of teams must be formulated, different models of team development must be studied, and the differentiation between individual and collective skills

and associated training strategies must be addressed. Specifying collective performance objectives must be undertaken before developing a measurement tool; thus guidelines are given for specifying conditions, tasks, and standards to develop satisfactory methods for collective training performance measurement. Chapter II also introduces the reader to several basic measurement concepts that must be considered in evaluating collective training techniques. The primary concern is for measurement reliability, including concerns for the performance itself, the observation of performance, and the recording of performance. Approaches to collective measurement need to consider that simulation techniques provide effective ways to include feedback, which is an essential component of training analysis. The chapter concludes with a summary of how preceding measurement issues, especially those concerning reliability, relate to Army Training and Evaluation Programs (ARTEP).

The third chapter presents a sample of the training devices and training settings used in the Army. Current military training approaches are discussed. Collective training approaches focus predominantly on simulation exercises and training devices to present training content.

The fourth chapter of this report elaborates on the concept of performance evaluation as it pertains to collective training systems described in the previous chapter. ARTEP ratings are fully described, including development of evaluation plans, training of evaluators, and preparation of test documents. Other evaluation systems used in the Army are reviewed. Many of these systems feature automated performance measurement, similar to Simulation Networking (SIMNET), with inherent shortcomings in after-action review, feedback, and human performance evaluation. Contrasted with computerized methodologies are prototype methodologies for the measurement of team performances, as exemplified by the measurement of team evolution and maturation (TEAM) as developed for Navy team training environments and by the Headquarters Effectiveness Assessment Tool (HEAT). These latter systems rely on observational data in conjunction with recording of data according to state-of-the-art rating criteria. There are advantages to studying both observational and computer-generated performance data simultaneously.

The fifth chapter readdresses basic psychometric issues involved in collective performance measurement. An expansion of reliability concerns focuses on three sources of measurement errors: (a) errors in the observation and recall of performance, (b) errors resulting from the instability of performance itself, and (c) errors in the recording of behavior due to deficiencies in the measurement instrument. Following general prescriptions and observations regarding accuracy and reliability, guidelines are related to a system of theorems contained in Wherry's classic

theory of ratings (in Landy & Farr, 1983). Finally, the complexities involved in integrating automated performance measurement (which focuses on outcome) with traditional observations of performance (which focuses on process) are related to new ideas in proxy and surrogate measurement.

The final chapter provides a summary of needs that includes: (1) further study of important dimensions of collective training; (2) utilization of critical incidents methodologies to identify fundamental characteristics of effective collective behaviors; and (3) development of reliable, standardized measurement systems over inherently variable conditions of combat. Suggestions for measurement improvement focus on the use of surrogate measurement systems to overcome unreliability in operational measures. A research program is suggested to comparatively assess subjective and objective measurement systems. The review concludes that computerized measurement data have not been adequately assessed in relation to alternate observational data or critical outcomes. Such research might determine a reduced set of optimal measures that might simultaneously relieve the trainer/evaluator's burden because of excessive information processing and provide the trainee with relevant information regarding his own performance. Better measures can also help training designers better determine retention of critical skills and the frequency of training reinforcement required for skill mastery.

Utilization of Findings:

These findings provide a research base for the future development of improved methods to assess unit combat effectiveness.

ASSESSMENT OF PERFORMANCE MEASUREMENT METHODOLOGIES
FOR COLLECTIVE MILITARY TRAINING

CONTENTS

CONTENTS (Continued)

CONTENTS (Continued)

ASSESSMENT OF PERFORMANCE MEASUREMENT METHODOLOGIES FOR COLLECTIVE MILITARY TRAINING

Chapter I. Introduction

This report describes the state of the art in team training methodologies in the military and provides recommendations for advancement in performance measurement techniques, particularly for the evaluation of collective training in military units. The study is part of an effort entitled "Unit Performance Assessment Research in Support of Readiness Exercises in Joint Services Interoperability Training" (Contract N61339-85-D-0024) undertaken by the University of Central Florida for the Training Performance Data Center and performed under the auspices of the Army Research Institute for the Behavioral and Social Sciences (ARI). The ultimate goal of this research study is to lay a foundation for future research and development (R&D) aimed at improving training performance measurement in the Army.

There are two points to be clarified before discussing potential improvements in training performance measurement methodologies in large units of personnel as exemplified by the Army. The first issue deals with the definition of "collective" as contrasted with traditional treatments of group performance. The second issue involves the distinction between the terms "measurement" and "evaluation."

In reference to the first of these points, "collective" performance assessment research has not been conducted extensively to date; most research has focussed on "teams" as the unit of measurement. Although much small group (team) training research may generalize to larger units (e.g., corps, division, brigade, battalion, platoon), the implicit assumption is that small group (team) research is more productive because it is "cleaner" from both a conceptual and measurement standpoint. For this reason, much of the research literature cited in this report deals exclusively with small group (team) behaviors that have been studied in controlled settings. Thus, it should be borne in mind that the terms "team" and "collective" are not synonymous. Army training focuses on individual versus collective entities, with collective training encompassing teams, squads, platoons, companies, and battalion levels. The location of training is at the institutional (schoolhouse) versus unit level. This report deals with collective performance at the unit level. It is important to determine whether team measurement methodologies are appropriate or feasible for the measurement of larger collective performances where personnel interactions and environmental variations are more complex.

The second point of clarification involves the distinction between the terms "measurement" and "evaluation." Measurement refers to the process of applying a set of metrics to an entity to provide a standardized basis of comparison among entities. Evaluation goes one step further by applying determinations of worth or value to such metric comparisons. "Assessment" is often used as a substitute for the term "evaluation". It is important to keep these distinctions in mind in reviewing the history of measurement and evaluation controversies within the Army and in attempting to develop appropriate methodologies for training performance measurement that may or may not contain evaluative components.

A. Background

The Defense Science Board Summer Study of 1982 and the 1985 Army Science Board Summer Study questioned the effectiveness with which the military services were measuring training performance. According to the 1985 study, the Army has several needs:

* "Quantitative" measures relating to training objectives, training strategies , and training effectiveness.

* "Quantifiable" tasks whose successful performance to standards leads to mission accomplishment.

* Evolution of programs to a "quantifiable" basis.

* Identification of task data needed to measure effectiveness of training.

* "ROI" (Return on Investment) information to guide expenditures of training resources.

* Knowledge of skills retention/learning rates to support unit sustainment training.

Since 1974, the backbone of the Army's unit training performance measurement system has been the Army Training and Evaluation Program (ARTEP). The ARTEP system was designed to apply criterion measurement to the field of collective training. It replaced the Army Training Programs (ATP) and Army Training Tests (ATT) which were relatively procedural and process-oriented rather than product-oriented. The ARTEP system was designed largely because General Creighton Abrams, Chief of Staff of the Army during the early 1970's, was dissatisfied with the current ATP and ATT training measurement

systems. Under these systems, numerical ratings which were assigned to units led to comparative evaluation and thus engendered a mentality on the part of field commanders to "pass the test" rather than to diagnose deficiencies (Report of Board of General Officers Appointed to Study Army Training Tests, 1959).

The ARTEP system was designed to provide a list of a unit's most combat-critical missions, accompanied by specific tasks, conditions, and standards, to accomplish required missions. For example, a task might be "use proper fire and maneuver techniques to eliminate opposing force resistance". The purpose of the ARTEP was to provide a strictly diagnostic device based on a YES-NO type of checklist so that immediate feedback would be available and corrections could be made. This type of system, with its focus on outcomes rather than the process of achieving outcomes, also minimized the threat of evaluative standards and the fear of failure that might accompany application of such evaluation.

However, Army leaders and training researchers soon uncovered several flaws in the ARTEP system. The problems can be summarized as follows:

* ARTEP manuals did not provide the training foundation necessary for units to attain proficiency in critical missions.

* Training objectives were vague and lacked specific standards of performance.

* No program existed for the integration of training from individual soldier to battalion level critical missions.

* Training methods and procedures differed from command to command.

The solution was identified as being one of providing commanders a training strategy to achieve proficiency for a unit's critical missions, describing a training plan that tied the "how to" with the "what to", supporting and enhancing standardization, emphasizing leader and unit proficiency, and providing a linkage between individual and collective tasks (U.S. Army Training Board Improved ARTEP Briefing, 1984).

In 1984, the Army's overall training strategy was revised to integrate individual and collective training through a building block series of individual and leader tasks, drills, and situational training exercises (STX's) which would lead to field training exercises (FTX's). According to this strategy,

both the institutional and unit training bases would share the responsibility for developing synchronized programs of instruction (POI's).

To implement this revision, an improved ARTEP system was proposed whereby mission training plans (MTP's) would be published at each echelon and would contain the training strategy for the echelon. Each would contain detailed training and evaluation outlines (T&EO's), leader training, STX's and FTX's oriented on critical missions, tests, and resource requirements. Complementing the "how to train" exercises, new training and evaluation outlines (T&EO's) were developed to provide a greater degree of objectivity. Since the MTP would be produced for each echelon, the T&EO's could concentrate on that echelon without producing an unmanageably thick document, as was the case with the former ARTEP system. New standards of performance were created to emphasize objectivity and standardization.

Despite the fact that the status of ARTEP is improving and the conversion from the dated ARTEP manuals to mission training plans continues, numerous problems remain in the Army's training evaluation system, especially as applied to collective performance measurement. The problems are not unique to the Army or to the military services in general, but relate to the inherent difficulty of assessing collective performance with a degree of reliability that enables appropriate use of training data. Examples of these problems will be discussed in Chapter IV, which details the ARTEP and other measurement systems.

B. Purpose

It is the purpose of this report to review the current state of military collective training performance measurement methodologies in order to define measurement problems and to offer guidelines for improvements in measurement systems. The topics of reliability, validity, and effectiveness are addressed in order to provide guidelines and to suggest areas for future research.

C. Approach

The approach to the literature review of current measurement systems and methodologies entailed a number of reviews of ARI publications as well as computerized literature searches of various data bases which would provide pertinent information.

The Army Research Institute List of Research Publications (1940-1983), provided titles of 1,332 research reports, 567 technical papers and reports, 582 research memoranda, 165 re-

search studies, investigations, and problem reviews, 155 technical reports, 188 research notes, and 82 research products. From these 3,071 titles, approximately 100 reports were ordered in microform and subsequently read for material relevant to unit performance measurement description. In addition, an updated listing of 573 ARI research publications (July 1983-July 1985) was similarly scanned and approximately 50 microform copies of relevant reports were ordered and subsequently reviewed and used in this research.

Technical report summaries from a computerized literature search of the Defense Technical Information Center (DTIC), using various key words, produced 264 abstracts. Few were found to be sufficiently current or appropriate for our purposes.

The final search for documents related to unit performance measurement methodologies involved a variety of non-defense files such as ABI/Inform, Dissertation Abstracts, ERIC, National Technical Information Service (NTIS), Psych Info, Psych Alert, and Sociological Abstracts. Although pertinent documents were found from most sources, the major source of relevant information was the NTIS data base. From a total of 478 abstracts covering the topics of team and crew performance measurement, 41 original reports were ordered and reviewed.

Because many of the research and technical reports were not current, an attempt was also made to locate information about newer training systems through the Army's Project Manager for Training Devices (PM TRADE), industry representatives, and researchers from universities and military laboratories.

All of these documents were reviewed in order to extract information specifically related to past, existing, and emerging training systems which focus on collective, rather than individual, performance.

Chapter II. Review of Team Training Research

This chapter provides a review of recent team training research to illustrate how performance measurement of teams and collective units will differ as a function of varying team training model assumptions, performance objectives, and measurement adequacy.

A great amount of effort has been expended from the year 1950 to the present concerning military team and related small-group research (e.g., Briggs & Johnson, 1965; Cooper, Shiflett, Korotkin, & Fleishman, 1984; Denson, 1981; Dyer, 1986; Glanzer & Glaser, 1955; Goldin & Thorndike, 1980, Hall & Rizzo, 1975; Kahan, Webb, Shavelson, & Stolzenberg, 1985; Meister, 1976; Morgan, Glickman, Woodard, Blaiwes, & Salas, 1986; Nieva, Fleishman, & Rieck, 1978; Roth, Hritz & McGill, 1984; Wagner, Hibbits, Rosenblatt, & Schultz, 1977). Despite a great deal of research interest, however, research progress is beset by a variety of theoretical and practical problems, particularly relating to the measurement technology available to support the training and performance of larger collectives (e.g., squad, platoon, company, etc.).

A. The Problem of Team Definition

In order for measurement improvement to occur, accepted definitions of team and collective must be formulated. The problem of team definition is not new (Briggs & Johnston, 1965; Denson, 1981; Dyer, 1984; Hall & Rizzo, 1975; Wagner et al., 1977). Dyer (1986) defines a team as consisting of two or more people with a common objective each having a specific duty to perform, where dependency is required for a successful completion of the desired goal. Morgan et al. (1986) say " a team is a distinguishable set of two or more individuals who interact interdependently and adaptively to achieve specified, shared, and valued objectives" (p. 4).

Hall and Rizzo (1975) list four characteristics used to define a tactical team (p. 9):

* It is goal or mission-oriented. That is, there is a specific objective for the team to achieve.

* It has a formal structure. For military teams, this structure is hierarchical in nature.

* Members have assigned roles and functions.

* Interaction is required among team members.

This report is mainly concerned with performance at the unit level. If units are defined as aggregates of many teams, a definition of team behaviors is necessary for the study and understanding of these larger units. Bauer (1981), in his review of methodologies for analysis of collective tasks, noted problems in defining what a collective task was in relation to individual tasks. For example, in defining what a collective task is, how do analytical scales developed for discriminating collective tasks relate to individual tasks? Conversely, how do individual tasks bear on collective tasks and the criticality of collective tasks? In spite of these problems, the aforementioned definitions of teams will be used throughout this report to apply to collective units as well, due to the fact that many characteristics of teams also apply to collective units even though their particular tasks may differ in some respects.

There are a great many studies that deal with team behaviors; however, there is very little written in regard to the theoretical base of team behavior (Dyer, 1986). In addition, according to Morgan et al. (1986), past studies have largely been conducted in laboratory settings and often neglected the consideration of operational constraints and requirements. Existing taxonomies of military teams focus on types of variables that must be measured but seldom provide direct translations to operational military teams (cf., Denson, 1981; Knerr, Root, & Word, 1979; Nieva, Fleishman, & Rieck, 1978).

Many forms of team training are possible, each having its own advantages and disadvantages. The following models illustrate the necessity to understand differences in teams when considering alternate forms of training and subsequent impacts on training measurement in operational contexts.

B. Models of Team Training

The "Stimulus-Response" model is an operant conditioning team training model most often used to train teams when situations are established. The problems and tasks can be specified in detail, and the response is limited to specific behaviors. Thus, rating the response is a checklist of behavior. This model is implicit in the ARTEP measurement system with its reliance on tasks, conditions, and standards. Adherence to or deviation from the specifications is most easily observed and recorded. The parameters of creative problem solving are most restrictive in this task situation. Since it is feasible to perform team training under established conditions in the laboratory, limited resources of time and money often call for

this type of situation. However, this model lacks the realism of most combat situations and relies on simple measurement as opposed to evaluation.

In the "organismic" model of team training, the emphasis is on viewing the team as a whole. Individuals are only sub-sets of the "organism." In this model emergent situations are stressed. While certain tasks are assigned, the solutions to the problems are more varied and open-ended. This leaves room for interpretation and creative problem solving by the team as a whole. Thus, it is more realistic for military applications since in battle, the solutions are not always clear-cut. The team must develop an effective team method of performance in order to deal with the ever-changing environment. Measurement of performance is therefore less objective, since behavior is not clearly specified and almost anything can occur. Recording of performance also calls for more interpretation of results, thus leading to evaluation as opposed to simple measurement. This evaluative component is also implicit in ARTEP. Team skills and compensatory behavior of team members are stressed in this condition. Because this situation is more common in real combat conditions, it is necessary to address this interpretive dimension when considering how best to assess performance under emerging task conditions.

Established and emergent situations are the context in which team performance occurs. These situations are the basis for the various team training models. Implicit in most team models is another dimension of concern, that of distinguishing between individual and team training.

C. Individual vs. Collective Training

It can be argued that team training is of no value unless the individual team members possess the minimum level of individual proficiency required for team proficiency. Many researchers concur that individual training and resulting proficiency should occur before team training (Daniels, Alden, Kanarick, Gray & Feuge, 1972; Finley, Heinlander, Thompson, & Sullivan, 1972; Johnson, 1966; Kahan, Webb, Shavelson, & Stoltzenberg, 1985; Kress & McGuire, 1979; O'Brien, Crum, Healy, Harris, & Osborn, 1978; Schrenk, Daniels, & Alden, 1969). This is generally due to the need for individual skills to reach stable levels of learning and retention before they can be expected to transfer to more complex situations involving teamwork.

A number of categories have been identified as predictors of group performance:

* Individual characteristics (general ability, task proficiency, and personality characteristics).

* Leadership (ability of the leader, personality, and leadership behavior).

* Group structural composition, or the mix of individual characteristics (general ability, task proficiency, personality, and cognitive style).

* Training techniques (feedback vs. no feedback, and feedback about group vs. individual performance).

According to Meister (1976), the training of individuals is time consuming and cost intensive. Yet, it is a mistake to give minimum attention to individual training and then expect the team training to take over and produce substantial results based on so called "on-the-job-training". Those individual predictors which relate to team performance should also be addressed in team training.

When examining team performance it is often difficult for the observer to know which is team performance and which is the result of an aggregate of individual performances. This ambiguity has a direct impact on performance measurement for, in order to improve performance, it is necessary to provide appropriate and accurate performance feedback. This requirement necessitates distinguishing between individual and collective performance (cf., Bauer, 1981) so that corrective behaviors can occur at the appropriate level.

A major distinction between team and individual performance environments is whether they are interactive or coactive. An interactive environment occurs when individual duties are collaborative and involve joint action, whereas coactive environments are those in which group productivity is a function of separate, but coordinated, individual effort. The majority of unit performance tasks in the Army are more interactive than they are coactive. The distinction between the two types is important because measurement and prediction of unit performance is dependent on whether the task is interactive or coactive.

D. Collective Skills

There are a great number of team skills relevant to combat training such as: communication, coordination, integration, self-evaluation, team awareness, and decision making. Yet most of these skills go greatly undefined, or no definition is accepted as universal. Thus, definitions are needed in order

for a more uniform understanding of team skills to be realized. To better understand these team skills, it is important to be able to isolate the effects of each team skill (Turney, Stanley, Cohen, & Greenberg, 1981) and focus on the interactions in the team skills in order to improve team performance. If these skills are isolated, then their effects can be measured and their pay off value determined (Turney et al., 1981). However, it is usually difficult to isolate skills because of interaction effects. This difficulty complicates the measurement process.

Training in team skills remains very important in emergent situation contexts, because it is in emergent situations that communication and coordination are essential in order to develop appropriate team procedures. Most military team training research therefore concentrates on C^2 (command and control) or C^3 (command, control, and communication) activities (cf., Cooper et al., 1984).

Communication, for example, may include such categories as: self-evaluation, cooperation, decision processing, problem solving, team awareness, pride, confidence, and aggressiveness (Kribs, Thurmond, & Dewey, 1977). Flowcharts are often used to analyze the work performance of the teams being studied (Dyer, 1986; Glanzer & Glaser, 1955; Thurmond & Kribs, 1978). These flowcharts are used to record communication, both verbal and nonverbal; team interactions; information channeling; decision making; computing; and supervising behaviors. Unfortunately these flowcharts are usually restricted to front-end analyses for the development of more training devices and programs (Dyer, 1986). However, Dyer (1986) suggests the use of flowcharts for analyzing the process that teams actually use during assigned missions.

Training in team communication skills has been shown to improve team performance (Turney & Cohen, 1981). Similar results could most likely occur in respect to other team skills, such as cooperation, team awareness, and decision making. Turney and Cohen (1981) also found that the content of messages is most often divided into the areas of commands and information (Dyer, 1986), and the frequency of occurrence of coordination and communications were major factors which helped to differentiate between better performing and worse performing teams. These results reinforce the importance of communication as a team skill, and imply that it is necessary to measure communication as accurately as possible. As we will see later, however, it is often difficult to measure communications at one location. It is necessary to be able to note the complex interactive aspects of this process.

Quality and timeliness of communication are important for scoring team and individual performance (Lahey & Slough, 1982). Lahey and Slough (1982) found, however, that quantitative measures of the volume of communication were of limited value. Their findings led to the conclusion that measures of quality of communication should cover correctness, completeness, and relevance in response to key events in an exercise or mission. These conclusions can be related to the fact that individuals have only a limited capacity for information-processing (Daniels et al., 1972). Communication takes up a portion of the information-processing capacity and thus limits the overall performance of the individual. Therefore, in assessing performance, perhaps only task relevant communication should be used in order to promote information-processing that will be task directed (Daniels et al., 1972). Experienced crews communicate less during routine missions and more during weapon delivery (Dyer, 1986; Obermayer & Vreuls, 1974; Obermayer, Vreuls, Muckler, Conway, & Fitzgerald, 1974). This result further supports the quality of communication theory as well as the overload theory. If the teams are able to work better with a minimal amount of communication, then perhaps it is due to the quality of communication that is taking place in the more experienced teams, as well as the fact that there is more time to concentrate on performance.

Thus, recent initiatives in the measurement of team and collective training, to be discussed in later chapters, have concentrated efforts in the development of headquarters command exercises and evaluation of other simulation techniques to train and measure these relevant team skills.

E. Specifying Collective Performance Objectives

One problem with collective performance measurement is that performance objectives are often not clearly described. Thus, it is difficult to interpret which behaviors are to be established in both training and testing. As we shall discuss later, this was a problem with early ARTEP's. Another problem is that, if objectives are clearly defined, they do not necessarily specify the conditions and standards under which the behaviors are to be established. Lack of performance conditions and standards specification is a serious detriment to team training and performance evaluation.

1. Specifying Tasks

Task analyses can provide a wealth of information on how to successfully perform a job or task function. Since even the most careful analyses may not be totally comprehensive, analysts who have conducted job task analyses often rely on

"experts" in the job area to cross-check their work for any omission of relevant data. Flanagan's (1954) Critical Incident Technique is often used. It compares effective task performance with ineffective performance in order to determine the critical incidents responsible for each.

Boldovici and Kraemer (1975), however, state that while task analyses may provide one with information on what needs to be done in order to perform effectively, it does not necessarily provide one with information on what needs to be taught. A learning analysis should be performed after the task analysis has been completed. Again the problem of selectivity arises, "What to include and what to exclude?" Since this is a subjective area, Flanagan's (1954) Critical Incident Technique may be used to improve inter-rater agreement. The goal of task analysis and learning analysis is to provide the trainee with instruction on critical skills and to provide the evaluators with an objective method to evaluate performance on these skills.

2. Specifying Conditions

One purpose for the specification of conditions to modify objectives is as a cross check for the evaluator to determine whether the objectives have been met. The conditions and objectives must therefore be stated in a specific, detailed, and preferably quantitative manner (Boldovici & Kraemer, 1975).

According to Boldvici and Kraemer (1975), it is impossible to list in the objective all the possible conditions that might occur in a task or mission. Determining the most important conditions, therefore, is most often a subjective task on the part of the writer of the objective. However, writer opinion may produce low reliability due to the fact that it is based on opinion of an "expert", and expert opinion may vary.

Once conditions are established, it is necessary to decide on the levels and modifiers that will need to be attached to these conditions. Theoretically, all job performance requirement conditions that might occur should be included in the objectives. It is important to provide a list of modifiers that will give the most comprehensive description of the job, but not become too cumbersome. Kraemer and Boldovici (1975) suggested including only those levels or conditions in objectives that require responses that are different from the response required by objectives with other levels. Thus, it is not always necessary to set different levels if the situation does not make a difference in the performance of a mission (i.e., firing fast or slow at a stationary target). Also, from a psychomet-

ric standpoint, ratings of performance are more reliable if the behavioral items to be rated are unique and discriminable.

3. Specifying Standards

Perhaps the most difficult part of specifying performance objectives is the development of realistic and objective standards against which to evaluate performance. A knowledge of system requirements is needed. However, detailed subject matter information about system requirements is not always readily available. In addition, according to Boldovici and Kraemer (1975), developers of job standards often overestimate their skills as trainers, the skill level of their recruits, and their ability to stretch funds, so that standards may be unrealistic.

Boldovici and Kraemer (1975) have suggested that job performance standards would be more effective if they were based on the normative performance of job incumbents. They would be more realistic and more objective than standards reached by arbitrary decisions. However, setting standards on the basis of normative performance assumes that there will be a sample of incumbents to base the standards on. This is not always the case. In addition, normative performance standards may not accurately reflect the system requirements. For example, in the military, if the skill level of the job incumbents is not better than that of the opponent, then the standards will not reflect the system requirements for a successful mission and may therefore not be relevant. The specification of performance standards is an extremely complex issue involving proper definition of multiple criteria, both subjective and objective. These issues will be addressed in Chapter V.

4. Difficulties in Specifying Objectives, Conditions, and Standards

In order to specify unit performance objectives, one must have an accurate idea of the tasks which are important to the mission and their criticality. These task statements should be accompanied by specific, detailed, and quantitative conditions which modify task statements and by standards which form a basis for evaluating task performance. The generation of clear, unambiguous objectives is easier said than done, however. Whereas task statements may be relatively easy to generate and verify objectively, sources of unreliability are introduced when conditions and standards are applied. Conditions may vary widely so only the most common will usually be incorporated in training objectives. This is a particular problem in emergent operational contexts.

Knerr et al. (1979) point out that tactical outcomes depend on several factors other than the proficiency of the units: interactions among force mix, missions, weather, and terrain can influence tactical results. "For example, weather interacts with force mixes, since poor visibility favors dismounted troops to the disadvantage of long-range weapons. If visibility improves during the actual action, then the advantage reverts to the long-range weapons. Because of such interactions, the outcome does not necessarily indicate the relative proficiency of the opposing forces. The impact of such external factors must be considered in evaluating the results of an exercise" (p. 3). In their study to develop process measurement in the armored cavalry, Knerr et al. (1979) found that, to support training objectives, the general requirements of the mission had to be stated very specifically, and observer activities had to be explicitly defined in order to develop a satisfactory method for unit performance measurement.

Combat effectiveness also depends on preparedness for the elements of surprise, and it is difficult to incorporate these conditions in training objectives. Similarly, standards of effectiveness may vary as a function of tasks and conditions. Thus, where complex team performances are required, it may be necessary to use a compensatory strategy, where certain standards on some tasks are relaxed in order to attain high levels of performance on others. In short, collective performance objectives should be specific enough to be trained and assessed but not so totally rigid as to inhibit intelligent interpretation of the overall combat picture.

F. Measuring Collective Performance

Most military operations depend on the integrated performances of teams of individuals who must coordinate activities to improve group decision making, unit performance, and operational effectiveness (Morgan et al., 1986). However, numerous authors (e.g.., Alluisi, 1977; Baum, Modrick & Hollingsworth, 1981; Denson, 1981, Dyer, 1986; Goldin & Thorndyke, 1980: Hall & Rizzo, 1975; Nieva et al., 1978) point to gaps in the analysis, definition, measurement, design, and evaluation of team training and performance.

There is no standard system or methodology for measuring unit/team performance. If one were to be developed, it would need to be quite extensive and general, with its coverage comprehensive in the area of measurement for all possible military occupational specialities. In addressing this problem, Wagner et al. (1977) developed a set of guidelines for a comprehensive team performance measurement system which included the follow-

ing:

* The definition of team performance objectives in terms of specified, observable outcomes to include criteria for acceptance and conditions of performance.

* The definition of a range of values applicable to each specified observable event.

* The detection, measurement, and recording of the value of an observed event at each occurrence.

* An evaluation of the team as having attained or not attained the objective--based on discrepancies between outcome criteria and observed event values.

* The feedback of team performance data to the training environment.

We have already discussed team performance objectives. This section covers the area of measurement and touches on the evaluation of measurement, its reliability, and validity. More on evaluation of units/teams will be given in Chapter V, which discusses in depth the major issues in performance measurement.

1. The Concept of Measurement

The lack of adequate performance measurement of personnel has plagued the military for years and has hindered research and development efforts in many areas. In 1952, Ericksen observed that large numbers of research problems directed at improving methods of training were becoming more and more severely bottlenecked by fundamental needs for improved methods of measuring proficiency. Since 1952, many authors have cited problems in present performance measurement systems.

The requirements for effective measures have been reported by many different authors (cf., Lane, 1986). Ruis, Spring and Atkinson (1971), for example, list the following items as requirements for good, individual, performance measures:

* Determining the present proficiency or capability of an individual.

* Predicting the future performance of an individual.

* Diagnosing individual strengths and weaknesses.

* Qualification for advancement or movement to a later stage of training (minimum standards/quality control).

* Feedback to student or instructor about progress.

* Evaluating alternative training methods.

These requirements address measurement on an individual scale. While many of the aspects of individual proficiency may be generalized to groups, teams, or units, there is still not a great amount of literature on team or collective performance measurement. Additionally, most of the studies conducted on performance measurement reflect the developments of the previous decade (Lane, 1986). Thus state-of-the-art reviews (such as Semple, Cotton & Sullivan, 1981; Vreuls & Obermayer, 1985) reflect the knowledge of the past decade and not necessarily the knowledge of the present state.

2. Measurement Reliability

An extremely important, but often overlooked, truism is that the quality of measurement tools most often determines the success of a research effort (Dyer, 1986). Without reliability, measurement is useless. Boldovici and Kraemer (1975) list two kinds of reliability that are of great importance:

* Inter-observer reliability: the extent to which two or more observers produce similar results in measurement.

* Stability: the extent to which measures taken at one time are representative or predictive of measures taken at another time (stable over time).

However, seldom are test psychometric evaluations performed. For example, in discussing test development for criterion-referenced tests used in Army testing and assessment centers, Swezey (1978) notes that, after test items are selected, Army test developers usually do not assess reliability and validity, at least in a strict psychometric sense. (Reliability, as noted above, refers to the stability of measurement across repeated testing or different observers; validity entails whether the test is a true measure of the behavior it purports to measure). Instead, a relatively informal process is used: the tests are administered several times and items that cause a great deal of difficulty are reviewed to see if they are constructed properly.

When Swezey surveyed Army testing and assessment centers and asked the question, "Do you use an item analysis technique?", 52% of the Army installation respondents (N=50) replied yes. Only 33% measured test reliability, and 26% computed coefficients of reliability. When asked, "Do you aid in validating tests?", 33% gave a positive response, but 36% reported using only content validity, the most elementary form of validity.

Considerably less than half of the test administrators at testing and assessment centers on four Army installations said that they were familiar with team performance testing situations. However, of those who indicated familiarity with the concept, many indicated that team performance testing is often individual evaluation in a team context. Actually, concluded Swezey (1976), the testing of team performance was very limited on the Army posts visited.

The assessment of reliability in operational (as opposed to school house) situations is more complex and, thus, with the exception of isolated studies (e.g., Olmstead, Baranick, & Elder, 1978), little evaluation of testing methods is done in the field.

It has been stated above that, without reliability, performance measurement is rendered useless. Yet neither inter-observer reliability nor stability has been reported in measures using various questionnaires, interviews, "expert judges", and rating scales, including ARTEP. This is unfortunate since so many military training measurement systems are based on these types of measures. It is also unfortunate that so much government military funding is going into the area of creative measurement while overlooking these two basic reliability issues (Boldovici & Kraemer, 1975). According to Boldovici and Kraemer, while the results of such research may be interesting, this research can not expand the knowledge base if it is based on unreliable measures.

In studying performance evaluation for Navy Anti-Air Warfare training, Chesler (1971) found that single performance measures were inadequate for evaluating system effectiveness and that a simple addition or lumping together of single measures did not improve the quality of system evaluation. It was therefore suggested (Wagner et al., 1977) that a deviation from linear models would be advantageous. Some type of "multiple cutoff" model would be most appropriate and would identify "compensatory" and "non-compensatory" activities. Compensatory behavior occurs when other team members are able to cover for the lack of skill or the poor behavior of other team members. Noncompensatory behavior is when the failure of one subset of

the team causes a detriment to the entire team performance or even failure of the mission. A multiple cutoff approach could be used to identify noncompensatory activities so that they would not interfere with system effectiveness at any stage. However, again, one would still require a reliable measurement system to identify such behaviors.

3. Sources of Measurement Reliability

The reliability of a measure, as stated before, is the most basic issue in proper measurement. If measures cannot be dependably replicated over time, then other criteria of measurement and quality are of little or no importance.

Unreliability can be caused by a number of different sources. Ronan and Prien (1971, cited in Lane, 1986) report two main sources of reliability: a) the reliability of the observation of the performance (this includes both objective and subjective measures) and b) the reliability of the performance itself (the activity actually being measured). In addition, the measurement instrument may cause unreliability. These three sources of measurement error will be discussed at greater depth in Chapter V, Issues in Performance Measurement. However, a brief overview of these issues, particularly as they relate to ARTEP, is presently in order.

Taking the first of these sources as a starting point, measurement has often been viewed as consisting of three phases: observer preparation, observation, recording and reporting. The manipulation of variables within these three phases can increase measurement reliability (Boldovici & Kraemer, 1975).

a. Observer preparation.

If the understanding and consistency of evaluators can be increased, then the reliability of their observations can probably be increased. However, Havron, Albert, McCullough, Johnson, and Wunshura (1978) found the following inadequacies in evaluator training in regard to the use of ARTEP's:

* Failure to delineate all evaluator functions and train in these.

* Little instruction as to how to use Training & Evaluation Outlines (T&EO's).

* Lack of instructions as to how to evaluate performance requiring coordination between elements not in visual contact.

* Lack of guidance for handling field critiques.

* Finally, a lack of appreciation of the complexity and difficulty of evaluator tasks as individuals, and as a team.

These are all supposedly problems that can be worked out in a reasonable fashion. Knerr et al. (1979) suggest that improved measurement can be achieved if the researcher:

* Specifies and defines concretely as possible the behaviors to be observed.

* Requires data collection personnel to observe, but not to judge the behavior.

* Trains observers fully.

* Requires observers to record observations immediately on clear, concise, easy-to use forms.

According to Knerr et al., if the above suggestions are followed in observer preparation, it is sure to increase the reliability of their observations beyond that of simple on-the-job-training for observers.

Boldovici and Kraemer (1975) recommend the following four steps to assist observers/evaluators in preparing for their task:

* Specificity of instructions. Reliability is likely to be greater when the instructions to observers are highly specific than when instructions are general and loosely stated.

* Timing of instructions. Instructions to observers should not be given so far in advance of observation as to permit forgetting and not so late as to preclude learning.

* Practice in observing and recording. Measurement reliability will be greater when observers have practiced measuring and recording the events of interest than when they have not. The practice variable interacts with timing of instructions, in that instructions to observers should be given far enough in advance of observation to allow time for practice.

* Testing observers. Measurement reliability can be indirectly increased by the use of tests given to observers to make sure that they are capable of performing whatever operations will be required of them.

Olmstead, Elder and Forsyth (1978) designed a study to measure the feasibility of training Organizational Effectiveness Staff Officers (OESO) and to provide feedback from their observations so that improved performance would result.

Olmstead and his colleagues concluded that "the ease with which the OESOs assimilated the concepts and recommended procedures and applied them within the context of an operational tactical operations center (TOC) leaves little doubt that properly qualified OESOs can be trained to assess process performance and give feedback on the results of such assessments in a meaningful manner" (1978, p. 15). This was demonstrated by a significant and strong positive relationship between the quality of battalion command group organizational process performance and combat outcomes as measured in battle simulation, when compared to trained and untrained OESO feedback.

The concepts of this study have considerable potential for improvement of combat readiness in operational military units. For example, the Army Training and Evaluation Program (ARTEP-1978) philosophy calls for a unit to be assessed on outcome variables, (e.g., mission accomplishment), and there is some provision for assessment of command group proficiency in performance of certain major tasks deemed relevant for accomplishment of the combat missions. "However, ARTEP (1978) assessment practices do not include diagnostic procedures that will enable commanders to better understand and assess the performance by battle staffs of the organizational processes that affect combat outcomes" (Olmstead et al., 1978, p. 4). If these diagnostic procedures could be developed, they would constitute an important addition to the training capabilities of unit commanders.

b. Observation.

At the present time, a variety of teams and team missions exist. For this reason, no standard procedures exist for observing teams and their missions (Dyer, 1986. However, reliability will always be affected by certain variables at work during the observation phase. This is true even in the case of highly standardized observers who have undergone extensive training and practice. As Boldovici and Kraemer (1975) have noted, reliability will decrease with the extent to which the

event observed is time-shared with other events or is not directly observable. These factors should be accounted for when examining the results of the observations.

Kubala (1978) posits that selection of the wrong Measure of Effectiveness (MOE), or excluding a critical MOE, will lead to wrong or poor conclusions about effectiveness. If this is true, and common sense suggests that it is, then further training on what to observe and record is necessary to avoid such problems.

There is an ongoing controversy as to the relative superiority of objective versus subjective measure of performance. Some analysts (Hagan, 1981; Kahan et al., 1985) believe that an objective measure of groups, based on small behavioral segments performed by working groups, will yield more reliable and valid results than subjective measures of performance.

One reason for the apparent superiority of objective ratings is that subjective ratings are biased by impressions of effort, rather than being pure measures of achievement (Hagan, 1981). Others (e.g., Lane, 1986) believe that subjective measures are often superior because many facets of performance can be integrated into an overall judgement.

Issues involved in observation of behavior are complex and varied. Also, assumptions regarding proper observational procedures influence measurement recording and reporting decisions. Thus, further discussion of observational issues will be discussed in more detail when we delve into fundamental reliability concerns in Chapter V.

c. Recording and reporting.

Because most of the recording and reporting of military training is done by personnel not trained as researchers (Dees, 1969; Dyer, 1986; Finley et al., 1972; Havron et al., 1955; Schrenk, Daniels & Alden, 1969), the reliability of the data could be questioned if evaluations are not properly prepared and their measurement tools are not as standardized and objective as possible. Reliability will again be affected during the recording and reporting phase. According to Boldovici and Kraemer (1975), the following variables affect reliability during the recording and reporting phase of measurement:

* Timing. Measurement reliability will increase with the decreased time between observation of the event of interest and recording of results.

* Design of recording forms. Well-designed data recording forms minimize the amount of judgement and decision-making required for their use and thereby increase the reliability of recorded results. Simplicity in data-recording forms, for example minimizes data-recording time, and therefore allows more time for observation.

Again, more will be said about the recording and reporting of observations in Chapter V. This chapter is designed to introduce the reader to basic reliability concepts in order to evaluate the advantages and disadvantages of various measurement systems used in the military.

G. Feedback/Knowledge of Results

In order for training to be effective, it is imperative to include feedback in the form of post-exercise critiques and reviews. According to Karanick et al. (1972), performance feedback is unquestionably the single most important parameter in team or individual training.

Feedback can be from two sources (Alexander & Cooperband, 1965).

* Intrinsic knowledge of results--feedback inherent in tasks.

* Extrinsic knowledge of results--feedback external from the system.

Both sources of feedback should be available in performance feedback models.

In order for feedback to be effective, Hall & Rizzo (1975) proposed that feedback needs to:

* Be immediate as opposed to delayed.

* Be instructive regarding what was done.

* Gradually reduce the amount of extrinsic elements so as to rely on intrinsic elements in the operational environment.

* Be direct as opposed to confounded with other factors.

Dyer (1986) warns that feedback that is too detailed during initial training may not be absorbed and may even be misused.

According to others (Hammell & Mara, 1970), immediate feedback seldom exists in the real world and is not always desirable. While it may exist in battle simulations, it is not likely to exist in a real battle situation. Thus reliance on immediate extrinsic feedback downplays the importance of intrinsic feedback, and may eventually deter performance when extrinsic feedback ceases to exist. This observation suggests that performance measurement should incorporate self-assessment as much as possible, with tasks set up in such a way as to minimize external immediate feedback.

A need exists in team training for both individual and group training feedback. For example, a study by Zander and Wolfe (1964, cited in Hagan, 1981; Kahan et al., 1985) found that group performance improved only when individual and group feedback was combined; neither group nor individual feedback was effective when given alone.

On the other hand, Nebeker, Dockstader, and Vickers (1975) compared the effects of individual and team performance feedback upon subsequent performance and found that any sort of feedback resulted in improved performance. Yet, increasing the amount or specificity of feedback had no additive effect and identification of individual team members did not enhance performance when the effects of feedback were controlled.

In addition, Lahey and Slough (1982) suggest that a multi-scoring system that evaluates team responses to significant events occurring in training exercises is needed for team evaluation and feedback.

Feedback is a major factor in augmenting and sustaining performance. Lack of feedback to the trainee is often associated with a lack of a sound assessment system for gathering information about trainee performance. Thus, to improve feedback and control training, valid performance assessment is vital.

H. Evaluation Objectives

According to Wagner et al. (1977), most often subjective or "armchair" evaluations are made in a checklist fashion; it is assumed that if the team completes the mission successfully, then it has received the proper training in the relevant behaviors. This is not always the case. Evaluation systems must therefore be evaluated themselves. Shaket, Saleh, and Freedy (1981) describe the following three objectives to be met for evaluating an evaluation system:

* Reduce the training and experience required of the training officer and staff needed to attain an acceptable level of training performance.

* Aid the officer to collect all the relevant and necessary data without depleting resources, manpower, etc.

* Reduce time needed for and improve the quality of the evaluation process.

Shaket et al. (1981) also describe how these general objectives can be met by providing assistance in the following more specific steps of the evaluation process:

* Aid the officer in identifying the relevant data and plan a collecting scheme to obtain the data (observers at the right point at the right time).

* Filter irrelevant data from the large amount collected.

* Concentrate the relevant data in usable form.

* Identify which tasks were accomplished and which were not.

* Identify missing tactical skills.

* Identify specific behaviors that caused eventual success or failure.

* Filter events that are peculiar to the particular exercise.

* Aid in focusing attention on major problem areas demonstrated in the exercise.

* Provide tutorial aids for any officer who is not familiar with a given mission, maneuver, or the particular tactical circumstance.

* Help to summarize and aggregate the problem areas identified so that most critical ones can be addressed first.

* Help to translate the deficiencies discovered into requirements for the next exercise to make it most effective.

These guidelines are extremely helpful in determining which evaluation systems are most adequate. However, in order for them to be an aid for updating existing evaluation systems (i.e., ARTEP) they would need to be expanded to more specific requirements.

I. ARTEP Reliability

As discussed before, the Army Training and Evaluation Programs (ARTEPs) were developed in order to correct deficiencies inherent in the Army Training Programs (ATPs) and Army Training Tests (ATTs). ARTEPs are to be used at the unit level for evaluating the performance of trained units. Since the outcome of these evaluations will be used to ultimately suggest and implement changes in the training process, it is necessary that the measurement process involved in ARTEP be valid and reliable. If this is not the case, or if the measurement is not comprehensive, it will result in inadequate or inefficient training.

According to Havron et al. (Vol. I, 1978), the ARTEPs have four main purposes:

* To evaluate the ability of a tactical unit to perform specified missions under simulated combat conditions.

* To provide a guide for training by specified mission standards or performance for combat-critical missions and tasks.

* To evaluate the effectiveness of past training of all echelons from crew/squad through battalion/task force.

* To provide an assessment of future training needs.

If the ARTEPs are successful in fulfilling their purposes, then they will emphasize: training, diagnosis of training deficiencies, performance, innovative problem solving, command performance, and reflect modification of tactics (Havron et al., Vol, I, 1978).

For ARTEPs to be successful, they must also comply with the sources of measurement reliability. To achieve reliability, evaluators involved in ARTEP evaluation using a battle simulation must be thoroughly and uniformly trained on the battle simulation to ensure that instructions to units, control techniques, and comparisons of performance with standards are consistent. The senior evaluator is usually responsible for

the development of evaluator training geared to the specific tasks elected for evaluation. According to an illustrative ARTEP Mission Training Plan (AMTP) for light infantry battalions (ARTEP FC7-13, 1985), in order to ensure properly qualified evaluators, the commander of the headquarters administering the ARTEP must:

* Issue the ARTEP Mission Training Plan to each evaluator and require thorough study.

* Review the ARTEP Mission Training Plan to clarify its contents.

* Conduct training on the particular battle simulation that will be used for the evaluation,

* Conduct an oral review of the battle simulation scenario, rules, and evaluation plan to develop courses of action

* Select a command group to demonstrate the play of the battle simulations to evaluate the command group's performance.

The evaluators must:

* Brief the command group to be evaluated on the evaluation concept and the battle simulation to be used.

* Conduct and control the evaluation according to the rules for play of the battle simulation and conditions outlined by the commander.

* Identify the critical points during the evaluation where command group task accomplishment can be objectively assessed on a "product" rather than "process" basis.

* Rate actual command group performance based on the standards prescribed in the ARTEP Mission Training Plan.

During the observer preparation phase, any number of factors may have occurred which would not have permitted the standardization of observers. Practice, time and money have all been cited (Boldovici & Kraemer, 1975) as common problems that may influence whether or not the observers are capable of doing a good job with high rates of inter-rater reliability.

During the observation phase, noise and communication overload influence the performance of observation. Too much may be going on for the observers to observe at once. Boldovici and Kraemer (1975) also suggest that measurement instruments may be to blame in this phase if they permit too much subjectivity. They also point out that strategies for measurement may be inappropriate, i.e., single observations may be used when the situation would best be suited to multiple observations over time or by multiple observers.

Finally, in the recording and reporting phase, time is of the essence. The more time elapsed between the observation and the recording of observations, the more inaccurate the objective recording will be. If recording is going on simultaneously with the observation phase, then an overload may occur, critical incidents will not be observed, and certain observations will not be recorded. Therefore, ARTEP, even the improved ARTEP, may be unreliable. Suffice it to say that the vast complexity of the ARTEP system serves as a psychometric challenge.

Some researchers have suggested that ARTEP could be computerized (Shaket et al., 1981) in order to reduce the time and communication loss due to flipping through the manual to find the related critical incidents. The Army Research Institute (ARI) recently funded a prototype of an electronic clipboard which would be used by ARTEP evaluators. This device would provide menus - selectable events for evaluation. Using a touch-sensitive screen, the evaluator would record the identification of the unit, his own identification, and whether a GO or NO GO assessment was made. At the end of the exercise or at some other appropriate time, the data stored in the clipboard could then be uploaded to a host computer for collation with other evaluation data and analysis (MacDiarmid, 1987).

Barber and Kaplan, (1979) suggest that performance evaluation could also be more reliable if the number of subtasks were reduced and if the standards could be more specific.

Because the Army has put so much emphasis into its ARTEP system, it should not overlook the problems associated with its reliability. Much research needs to be conducted on how to improve the reliability of the ARTEP measures, and we will suggest an innovative method to improve reliability in Chapter VI of this report, Research Implications.

J. Summary

In this chapter, we have attempted to cover some of the fundamental issues of team training to illustrate how performance measurement of teams and larger collective units will differ as a function of varying team training models and assumptions. Fundamental to the development of appropriate measurement methodologies for collective performance is the distinction between individual and group performance. Even when group or collective performance is specified, it may be difficult to isolate the individual and team skills which relate to effective performance. For example, although coordination and communication activities are believed to be essential components of effective team behavior, the interactive aspects of these activities make it difficult to apply uniform standards to performance. In effect, what is appropriate under one scenario may be completely inappropriate given a slight change in situation or condition (task, group composition, experience level, etc.). This difficulty greatly hampers advancement in collective performance measurement. Although simulation techniques offer one solution to this complex problem, other problems exist.

Dyer (1986) states that the limited progress in describing how teams function reflects the lack of techniques that have been actually developed and the diversity among the few techniques that have been used. No single descriptive procedure has been employed extensively or has been widely accepted by the research community. Similarly, Morgan et al. (1986) state that one of the main ingredients missing in team training is an objective, standardized, and practical measurement technology. Existing combat unit performance measurement techniques depend largely on judgmental data and often do not evaluate the unit's ability in the field (Hayes, Davis, Hayes, Abolfathi, Harvey, & Kenyon 1977; Knerr et al., 1979).

In addition, measures are most often developed for variables that are easy to evaluate (Dyer, 1986). This means that evaluated variables are not necessarily ones that are critical to a mission or ones that are reliably assessed. However, performance on any given task may have a unique effect on the total outcome of a mission (Connelly et al., 1980) regardless of whether or not the task is evaluated. Thus, there is always a question as to whether measures are sufficiently comprehensive. It has been suggested (Hackman, 1982) that a single methodology for effectiveness assessment will not be comprehensive. For example, Boldovici and Kraemer (1975) compared critical incidents with task descriptions in their critical incident study of reports by veterans of armored combat. Task descriptions were found to describe more elemental types of behavior, while

safety and novel solutions to combat and non-combat problems were most often found using critical incidents. Therefore, Boldovici and Kraemer concluded that task descriptions should be used for mastery of the basics, while critical incidents may be used for advanced training and end-of-course evaluation. Efforts should be expended to make both types of evaluation as reliable as possible.

Presently, the most common measurement approach is subjective evaluation (Hall & Rizzo, 1975). This approach, typified by ARTEP, usually makes use of satisfactory/unsatisfactory scales, thus reducing the evaluation process to a checklist procedure with little room for interpretation. Subjects receive a weighted score, with the team score usually being a summation of weighted individual scores. Introduction of automated performance measurement and other more objective methodologies may provide more comprehensive assessment.

In summary, more research needs to be done on how to improve the reliability of measures used in military contexts (e.g., ARTEP). Many researchers (e.g., Wagner et al., 1977) suggest a systems approach to team training to assist in the establishment of criteria/standards, while others (e.g., Kristiansen & Witmer, 1981) have developed a systems approach to "fix" training problems. Systems approaches, which include analyses, design, development, implementation, and evaluation of training programs, provide a way to assure training goals and thus rely on adequate system measurement to make proper adjustments to the system.

The guidelines in this section on measurement should be kept in mind as we turn now to a closer look at some of the current systems used in collective training in the Army.

Chapter III. Current Army Collective Training Approaches

The purpose of this chapter is to present a sample of the training systems used by the Army. This review is not comprehensive in coverage; many training approaches have become obsolete and are therefore not included herein. An attempt has been made to present those approaches that were felt to be most important, with an emphasis on those that were considered emergent for task situations as they present a greater level of realism. The review primarily addresses Army training approaches because Navy, Air Force, and Marine approaches are either quite similar or do not address training factors that are unique to the Army.

The Army has three categories of training: individual training, unit training, and training support (i.e., support personnel). Unit or team training takes precedence over individual training due to the nature of the Army's force structure (Rosenblum, 1979). According to Wagner et al., (1977), this team training is largely conducted in the field with no formal institutional training associated with unit training.

Army training settings have two classes: field exercises (ARTEP missions and engagement simulations) and simulations. Havron et al. (Vol. V, 1979) have described the field exercises as being mainly designed for entire units, while simulations are the superior method for training battalion and company commanders and staff.

A. Army Training and Evaluation Program (ARTEP)

The Army Training and Evaluation Programs (ARTEPs), with their Training and Evaluation Outlines (T&EOs), guide the Army's collective training. The ARTEPs also serve as a base for conducting most of the Army's field exercises and simulations.

ARTEP Mission Training Plans (MTPs) list keys to successful training. For example, it is stated (ARTEP, 1985) that the exercise commanders are responsible for the training of all individuals and units under their control. The emphasis in the training exercises is usually placed on the individual soldier, and battalion commanders must ensure that the platoon leaders realize the importance of the individual tasks. ARTEP stresses centralized planning and decentralized execution. The battalion

commander sets the overall training goals and objectives, and the subordinate commander develops the training plan for his unit based on the overall training goals. In order to be proficient, the platoons must be able to perform the mission assigned to them. Therefore, success of a battalion mission becomes a function of battalion headquarters proficiency in command, control, communications, and planning.

According to ARTEP instructions (ARTEP, 1985), the battalion commander must give adequate time for the individuals, squads, and platoons to become proficient at working together before assigning specific missions related to the accomplishment of a company mission. The commander assigns training drills that will result in the repetitive conduct of all the essential drills. This procedure also helps prepare the individuals and platoons for the successful accomplishment of their specific missions.

ARTEP training also emphasizes realistic training conditions, with night training being common. The training must be as realistic as possible in order to properly train and motivate the soldier.

The commander asks three questions throughout the training process: Where should we be? Where are we now? How can we best get from where we are to where we should be? "Where should we be" is answered within the pages of the ARTEP Mission Training Plan. The commanders may add tasks or make the tasks more difficult, but they can not make the tasks easier. The standards for appropriate performance are also provided in the ARTEP plan. "Where are we now" is answered by the most recent measure of performance, as determined by observation or evaluation in relation to minimum combat essential critical tasks and standards listed in the ARTEP MTP. After answering the first two questions, the commander can develop a training plan to answer the third question: "How can we best get from where we are to where we should be?" This new training program should overcome any weaknesses discovered in answering the second question.

In order to practice mission command and control properly, leaders and staffs are trained as teams. Usually additional training, other than the training received through daily routines, is required. Many effective techniques are available to the commanders for performance of additional training. These include on-the-job coaching and critiquing by senior leaders, tactical exercises without troops (TEWT), map exercises (MAPEX), situation training exercises (STX), battle simulations, command field exercises (CFX), command post exercises (CPX), and fire coordination exercises (FCX). TEWT and MAPEX

function with only the leader and selected subordinates. They consider mission-unique factors and how best to use the given terrain and tactics. TEWT is conducted in the field, whereas MAPEX can include the study of maps and the viewing of terrain slides. CFX, CPX, and FCX provide realistic practice on their mission while keeping troop and unit support to a minimum. They allow participants to practice command, control, and communication while under simulated wartime pressures. The various exercises available are listed below along with a brief description of when they are appropriate to use (ARTEP, 1985).

EXERCISES AVAILABLE

Map Exercise (MAPEX)	- Low cost; low overhead; multi-echelon; suitable for command and control at all levels; limited hands-on training.
Tactical Exercise Without Troops (TEWT)	- Low cost; low overhead; valuable for company and below; "terrain walks" at higher levels; valuable for contingency planning.
Simulation Training Exercise (STX)	- Low cost; low overhead; valuable staff and leader training reinforces collective training.
Fire Coordination Exercise (FCX)	- Medium cost; medium overhead; exercises total-unit capabilities in scaled-down live fire mode; can exercise leaders at all levels in the orchestration of weapons and support systems; suitable for platoon, company, and battalion echelons.
Live Fire Exercises (LFX)	- High cost; high overhead; resource intensive; generally limited to platoon and company level.
Combined Arms Live Fire Exercise (CALIFAX)	- High cost; high overhead; resource intensive; combined arms training, battalion level.
Command Post Exercise (CPX)	- Medium cost; medium overhead; exercises command, staff head

quarters, and communication personnel; suitable for battalion through corps; does not fully exercise service support.

Command Field Exercise (CFX) - Medium cost; high overhead; exercise total unit capabilities (weapons and support systems); suitable for battalion through corps.

Field Training Exercise (FTX) - High cost; high overhead; most realistic at lower echelons; requires extensive control measures; not suitable as a basic training method for staff.

Joint Training Exercise (JTX) - High cost; high overhead; exercise two or more services; more an environment than a technique; essential for combining all resources to fight on the modern battlefield.

Emergency Deployment Readiness Exercise(EDRE) - Medium cost; high overhead; exercises deployment capabilities; involves installation support activities in their service and support roles; normally conducted for the battalion task force level.

Battle Simulation Exercises - Low cost; low overhead; appropriate for training leaders and battle staff (ARTEP, 1985).

The training of the battalion headquarters is similar to training squads and platoons in that individuals (officer and enlisted) must be proficient in certain individual tasks before any type of team training can begin. After individual training has been completed, a road map of progressive training is followed which ultimately leads to battalion-level ARTEP. A building block approach guides this collective training. Each staff section is given specific training tasks which can be used in garrison. These training tasks help to develop the internal procedures for that section in field operations. Once these specific training tasks are completed, the staff begins to train as a unit and the training becomes progressively more difficult. Soldiers and NCOs perform a series of tactical operations center (TOC) drills, such as the tasks relevant to

setting up and moving the command post. The officers and senior NCOs participate in a series of staff coordination exercises. These exercises are used to increase proficiency in procedures relevant to combat.

Once the headquarters is trained and company-level proficiency has been attained, then battalion-level collective training in an FTX can begin. The purpose of the FTX is to verify proficiency in the training conducted to date and enhance unit readiness. The systematic training of the staff and increased levels of proficiency prepares the battalion for an external evaluation, by the National Training Center, REFORGER, TEAM SPIRIT, or other battalion-level exercises (ARTEP, 1985).

According to many researchers, one of the best ways to assess C^2 and C^3 activities is through simulation techniques because of their potential for controlling the training environment in order to simplify the measurement of complex coordinated team behaviors (Cooper et al., 1984; Knerr et al., 1979; Shiflett, Eisner, Price, & Schemmer, 1982).

According to Sulzen (1980), engagement simulation techniques are superior to operational or field training techniques such as ARTEP and also provide a more objective method for evaluating team performance. Evaluators in engagement simulation exercises classify the various tasks performed by the units and provide numerical, objective ratings instead of summary evaluations.

Engagement simulation techniques also simulate the complexities of actual battle. They provide training using opposing forces in mock battle situations. Casualties and/or hits are determined in a number of ways using telescopes, [illegible], numbers on targets, etc. Engagement simulation is very realistic, therefore, in that the situation is emergent. Often Training and Evaluation Outlines (T&EO's) are not at hand for finding the prescribed answers. The response must be creative and intelligent. Casualty assessment occurs in these two-sided, creative problem solving tactical field exercises through the simulation of weapons and weapons effects. Casualties are brought about in a unique and timely manner as a result of interactions of opposing forces.

Wagner et al. (1977) and others have found that engagement simulation techniques are valid for training groups. Soldiers and units alike reach higher levels of tactical proficiency more rapidly. Motivation is also high due to the realistic and challenging nature of the tactical exercises.

As we will discuss later, engagement simulation is best

when used as an objective tool for measuring casualty and damage outcomes. It may also be used for post-exercise critiques, after-action reviews, and for peer and self-assessments. Thus, engagement simulation combines the rigor of quantitative measurement with the simplicity and understandability of more qualitative measurement and feedback.

The major Army simulation training strategies and training settings are described below.

B. Squad Combat Operations Exercises, Simulated (SCOPES)

SCOPES (Squad Combat Operations Exercises, Simulated) are used for infantry squad training and fall under the heading of engagement simulations. Originally developed for the Army to teach movement techniques to M16A1 rifle teams and squads to increase proficiency, the SCOPES technique provides a realistic framework for battle simulation because it simulates live fire. "Hits" can be easily determined. There are real outcomes with "winners" and "losers". The outcomes can then be discussed in an After-Action Review (AAR). The reasons for success and failure are discussed, with suggestions given for improving ways in which to avoid being killed and to better attack the enemy.

C. REALTRAIN

REALTRAIN is another engagement simulation technique. It is more advanced than SCOPES in that it allows for the simulation of larger scale exercises using more than just the use of rifles. Soldiers who have trained under REALTRAIN as opposed to more conventional methods have performed better under the REALTRAIN conditions (Knerr et al., 1979), obtained higher levels of proficiency, and been more motivated to train (Wagner et al., 1977).

As technology advances, training systems become outdated. REALTRAIN has been no exception to this. SCOPES was replaced by REALTRAIN, and REALTRAIN has all but been replaced by the technology and advancement of the MILES system.

D. Multiple-Integrated Laser Engagement System (MILES)

MILES is a type of engagement simulation with the usual two-sided, free-play exercises. In MILES, casualties are determined automatically by eye-safe lasers. The casualties are reported to the Net Control Station (NCS) where a total is kept and the outcome is determined (Havron et al., Vol. V, 1979). This system allows for battle simulation in day or night. Its ability to report kill scores automatically reduces

the subjectivity in battle simulation exercises. MILES also allows for simulation in battalion and task force size teams (Wagner et al., 1977; Havron et al., Vol. VI, 1979). Current work is underway to improve MILES to operate under obscuration conditions (Industry Day, 1987).

Using lasers does have a good potential for reducing the amount of unreliable human measures (Wagner et al., 1977). According to Knerr et al. (1979), however, MILES, in reducing the need for human controllers, also reduces the amount of tactical activities data available. This reduction of available tactical data regarding performance may affect discussions immediately following the exercise, where controllers get together to compare results and validate the data. This is followed by an AAR (After-Action Review) in which trainees learn about their mistakes and their effective behavior. The AAR should provide knowledge and effective feedback to promote performance effectiveness.

The preceding simulations fall into the category of engagement simulations. The next group of simulations can be categorized as field exercises comprising battalion-level manual and computer-support/assisted simulations.

E. Combined Arms Tactical Training Simulator (CATTS)

The Combined Arms Tactical Training Simulator (CATTS) is an example of a system used in a field training setting. CATTS is a computer driven exercise (Wagner et al., 1977) for training maneuver battalion commanders and cavalry squadron command groups (Barber & Kaplan, 1979) to attain and sustain ARTEP standards in the control and coordination of combined-arms operations. The computer is located at Fort Leavenworth, Kansas, where it simulates the behavior of platoons/units in combat. CATTS maintains information on location of units/personnel; unit strength, equipment, ammunition, and other assets; fuel for friendly and enemy forces; movement rate depending on terrain/ weather conditions; etc. (Kaplan & Barber, 1979; Havron et al., Vol. V, 1979). The computer simulates the behavior of units in combat providing for a realistic free-play exercise that is conducted in a real-time mode (Kaplan & Barber, 1979).

Until Simulation Networking (SIMNET), the CATTS system was the most sophisticated simulation training device available for the battalion staff in the Army inventory (Havron et al., Vol. V, 1979). However, the situations of battle are limited to those that are programmed into the computer. While some limitations do exist, most of the ARTEP subtasks can be simulated in CATTS (Kaplan & Barber, 1979), thus demonstrating

the utility for training battalion command groups under the CATTS system.

F. Computer-Assisted MAP Maneuver System (CAMMS)

The Computer-Assisted Map Maneuver System (CAMMS) is somewhat similar to CATTS in that it, too, is a computer-assisted battle simulator for battalion commanders and their staffs. Symbols have been developed for units and weapons, these symbols are then put to use on a terrain board which simulates the actual battlefield. The automatic recording of movement, engagement, and administrative procedures helps to reduce processing demands of the controllers and adds to the value of the training (Havron et al., Vol. V., 1979)

However, CAMMS, like CATTS, is limited to those situations that have been programmed into the computer (Havron et al., Vol. V, 1979). CAMMS, unlike CATTS, is a portable system. A travelling team conducts the controller training, thus local unit training managers are responsible for selection of controllers and auxiliary personnel. This means that the CAMMS controllers and auxiliary personnel have not achieved the level of "expertise" that has been achieved by the more permanent CATTS controllers and auxiliary personnel. While this is a drawback of CAMMS, it is only a slight one. Both CAMMS and CATTS are highly recommended for battalion staff training (Havron et al., Vol. V, 1979).

G. Pegasus - GTA 71-2-1

Pegasus is a command group simulation, like CATTS and CAMMS. It is also for battalion and higher level command/staff members. Pegasus, unlike CATTS and CAMMS, is not computerized, it is entirely manual. The controllers are responsible for executing casualty assessment, orders, maintenance of strengths, status, etc. (Havron et al., 1979). These are the areas that are normally handled by the computer in CATTS or CAMMS. CATTS, CAMMS, and Pegasus were all developed to represent command group tasks outlined in ARTEPs 100-1 and 71-2. According to Havron et al. (1979)), their use (especially CATTS) pointed out three areas of weaknesses in ARTEPs: performance of subtasks, specific performance deficiencies, and the relative contribution of individual subtasks to overall effectiveness. Because Pegasus is entirely manual, controllers have to spend more time processing information and data. This processing means that less time is available for filtering the information (i.e., observing player performance) and acting out their role as controllers. Pegasus was also designed to be interactive and to give the outcomes of player behavior. But,

because the system is manual, an overload of processing may occur, thus causing a lag in the real time of event occurrence. A system overload may also cause administrative/logistical errors to occur. They will simply be overlooked due to the necessary extra processing. Pegasus does have the most comprehensive capabilities for evaluating staff performance diagnostically. However, the above mentioned problems are likely to occur due to an overload of trying to process this comprehensive set of materials. Thus, Pegasus, while not being limited by any computer programming deficiencies, is limited by the information processing capabilities of the controllers and also by
their manipulation abilities.

H. Military Airlift Center Europe (MACE)

MACE is a microcomputer-based battalion/squadron training simulation using Europe as its scenario. Its purpose is also to provide a low cost computer based training environment for commanders and their staff in command and control to attain and sustain ARTEP standards (ARTEP, 1985).

I. Dunn-Kempf

The Dunn-Kempf tactical board game (a nonengagement simulation training device) trains command and control leaders at the company level and platoon leaders in exercising tactics and maneuvers. Dunn-Kempf uses OPFOR (Opposing Forces) tactics, but does not incorporate the effects of individual weapons. A board is used to represent the terrain. While this is certainly convenient, it is also unrealistic in that it allows the commander/platoon leader to see all of his forces. This is usually impossible in the actual battlefield, due to any number of constraining factors (e.g., distances, topography, vegetation, etc.).

The system is manual, thus all battle pertinent information must be recorded by the controllers. Here again, the problem of information-processing overload is likely to occur, due to the information processing and manipulation limitations of the controllers. This affects the realism of the simulation. Another criticism of Dunn-Kempf is that it occurs at the company and platoon level. Havron et al. (Vol. V, 1979) believe that a transfer of learning is less likely at this level than at the battalion staff level, which are more realistic to combat jobs.

J. Small Combat Unit Evaluation (SCUE)

SCUE, the Small Combat Unit Evaluation game, is a manual system for simulating field exercises. SCUE was developed by the United States Army Research Institute (ARI). It, like Dunn-Kempf, is also for company, platoon, squad, and fire teams. Thus, since it is manual and for the same unit echelons, it faces many of the same limitations as Dunn-Kempf. SCUE does, however, incorporate the use of individual weapons (i.e., M-16), but does not use OPFOR tactics (Havron et al., 1979).

K. Joint Exercise Support System (JESS)

JESS and the following systems represent more current computerized battle simulations which were designed to depict higher-echelon engagement scenarios. JESS is a two-sided, stochastic, ground force model designed to train corps and corps major subordinate command (MSC) commanders and their staffs in the proper implementation of Airland Battle Doctrine. JESS is being developed as a model for Corps Battle Simulation (CBS) by Jet Propulsion Laboratory (JPL) in Pasadena, California. The model, presented on screen by hexes to controllers, represents the terrain by videodisc scenery overlayed by computer-generated graphics. The system depicts units from corps down to battalion for its CBS applications. However, company-size forces can be created and played for blue (friendly) forces. Red (threat) forces are normally shown down to the regimental size.

Maneuver units can perform different missions (attack, defend, withdraw, move, delay, etc.). Their position, combat strength and weapons range, by weapons system, is described by the computer, as are combat support units, such as field artillery, Tactical Air (TACAIR), attack helicopters, cavalry, air defense, engineer, and NBC units. Combat service support units are also portrayed.

The players communicate with controllers via the voice and teletype capabilities of their organic signal units. Information is sent to controllers as orders to lower level commanders, while reports from lower level commanders are sent to the player by the controllers.

JESS has checkpoint, restart and replay capabilities, and it runs in real time. (Battle Simulation Software Survey, 1986).

L. JANUS/JANUS (T)

The JANUS system is a two-sided stochastic, high-intensity combat model developed by the Lawrence Livermore National Laboratory to train brigade and battalion commanders and staffs in ground combat doctrine. The Training and Doctrine Command (TRADOC) has upgraded the system in JANUS (T) to improve the training aspects.

JANUS provides a simulation model depicting digitized terrain on the monitors, where troop/unit positions and movements are based on the actual terrain. The units are portrayed as battalions and companies normally but can be broken down into individual weapons systems (e.g., tank, APC, M-16). Threat units are played from red division down to battalion. Maneuver units are tracked by type, strength, and position. Line of sight calculations determine what targets can be seen and/or fired upon. Units can perform any normal ground maneuver. JANUS also provides field artillery, air defense, artillery and attack helicopter play, although it lacks intelligence modeling, TACAIR, and other features.

The commander and his staff use organic communications equipment to communicate with lower level commanders and staff who are located with the controllers and provide inputs verbally. They receive information visually from video display units. The controller is usually one of the lower level commanders who has received 4-6 hours of training in using the workstation equipment (Battle Simulation Software Survey, 1986).

M. Army Training Battle Simulation System (ARTBASS)

ARTBASS is a mobile, two-sided, mainly stochastic combat simulator developed by Singer-Link and designed for battalion and brigade levels. Although maintained by the U.S. Army Combined Arms Center at Ft. Leavenworth, Kansas, it is currently being fielded for general use throughout the Army. The Combined Arms Tactical Training Simulator (CATTS) described earlier, provided the test-bed for ARTBASS. Designed to provide training via realtime battle simulation, ARTBASS has digitized terrain representation which can play any type of terrain. ARTBASS depicts units from brigade down to the squad level with units portrayed as aggregate of weapons. It provides combat resolution at the combat level and portrays red units down to the battalion level. For blue forces, all combat arms are portrayed as well as combat support units (field artillery, TACAIR, air defense artillery, electronic warfare, aviation, NBC, and helicopters) and combat service support units.

The players coordinate with the controllers via the normal communication channels of the organization. Controllers act as company commanders, fire direction centers, logistic support facilities and other combat support and combat service support elements. ARTBASS also has checkpoint, restart and replay capabilities (Battle Simulation Software Survey, 1986).

N. Automated Support System for Army Unit Logistic Training (ASSAULT)

ASSAULT is a multi-echelon logistics exercise training simulation being developed by BDM to provide combat service support (CSS) to unit commanders in a training situation and to test the feasibility of simulation-supported CSS training exercises (Battle Simulation Software Survey, 1986).

O. Computerized Battle Simulation (COMBAT-SIM)

COMBAT-SIM is a two sided battalion combat simulator, originally developed for the Australian Army War Game Center by Perceptronics, and used to train battalion level commanders and their staffs in decision-making processes in simulated combat environments.

Area maps are generated from video disks and overlayed with computer-generated imagery for representation of units, movement, obstacles, and combat fires. COMBAT-SIM represents units as aggregates of weapons systems. The system normally utilizes units of company and platoon size, while threat units are portrayed at company level.

COMBAT-SIM portrays all types of maneuver units with the capability and maneuverability of each unit determined by the current "operational state" of the unit. The "operational state" combines descriptions of the tactical posture of the unit, its maximum rate of movement, firepower effectiveness, and vulnerability to observation and fire. COMBAT-SIM also portrays field artillery fires, air defense fires, and TACAIR functions. Error probability, round distribution, probability of hit and of kill are used to determine results. Combat service support units are also modeled.

Training audience/role player interaction is provided by organic unit communication equipment, where the training audience uses its standard operating procedures for coordinating with the role players. The system checkpoint, restart and replay capabilities uses archival data. The archival data are also used to produce postgame summary reports. A controller can be trained to use the system in 4 to 6 hours. (Battle

Simulation Software Survey, 1986).

P. Battlefield Management System (BMS)

The Armor Center at Fort Knox (Pappa, 1986) identified needs for a system which integrates all the various combat systems found on the close combat heavy battlefield at the battalion and below. To meet these needs, the Battlefield Management System (BMS) has been proposed to integrate all of the previous battlefield methodologies. As currently envisioned, the system will contain a common data base that all subsystems will share to allow the commander to optimize his weapons against any threat. It is this data base that will enable much of the decision-making process and staff procedures to become embedded in software that will cue the commanders to critical situations on the battlefield.

The BMS is being developed in order to correct three deficiencies that were identified in the areas of target acquisition, communications and command, and control.

The Battlefield Management System is an electronic information gathering, processing and distribution system. The system is designed to quickly gather and process real time battlefield information and intelligence in a responsive manner.

The Army has developed the Army Command and Control Master Plan and Army Battlefield Interface Concept which are the overall capstone documents for the integration of horizontal and vertical information. Maneuver control, intelligence/electronic warfare, fire support, combat service support, and air defense are the elements of this C^2 master plan. These nodes and their relation to one another are known as the Sigma Star Concept.

These systems were developed due to problems associated with lack of time available to effectively coordinate combat, combat support, and information gathering systems. The technology of the BMS should help commanders have a more accurate overall picture of the battlefield which should enable them to fight more efficiently and execute faster than the enemy. "In broad terms the Battlefield Management System substitutes technology for time consuming manual functions at the tactical level, and expedites surveillance and fire distribution activities" (Pappa, p. 12).

The major inputs into the BMS will be sensor information, command and control information to include orders and status of forces, targeting data for both direct and indirect fire, and navigation and logistics data. The Battlefield Management Sys-

tem will ultimately include all combat systems, armor, infantry, artillery, aviation, and air defense at the levels of command up to battalion and will attempt to use embedded training devices and system redundancy.

1. Simulation Networking (SIMNET)

As part of the BMS concept implementation, Simulation Networking, or SIMNET, was developed as an advanced high-risk technology project sponsored by the Defense Advanced Research Projects Agency (DARPA) in close cooperation with the Army (Herzog, 1986). Its objective is to develop the Department of Defense technology base for the large scale networking of military training simulators. This technology is designed to permit regular and intensive practice of combat skills by large-teams, and is viewed as an essential technology for the future preparedness of U.S. combat forces.

The SIMNET project began in 1983 when analysts from DARPA analyzed the Army Training and Evaluation Program (ARTEP) and determined the tasks and skills that are required to successfully complete the mission. Primary ARTEP missions include:

* Movement to contact and the hasty attack
* Deliberate attack
* Hasty defense
* Deliberate defense
* Passage of lines (forward and rearward passages)

The analysis included a review of the major missions, the collective tasks that support them and the network of supporting common skills and individual tasks that assure accomplishment of the collective tasks. All of the individual tasks were evaluated as to their potential feasibility for simulation. DARPA focused its technological development on armor (the M1 tank) and mechanized infantry (the M2/M3 Bradley Fighting Vehicle) training requirements. Another major issue of SIMNET is that of local area networking (LAN) and its ability to support a cluster of simulators at a single site. The number of simulators may vary from as few as 2 to as many as a 100. An example of the LAN system is the cluster scheduled for the U.S. Army Armor Center at Fort Knox which will have 82 simulators simulating a tank heavy battalion task force featuring M1 tanks, M2 and M3 Bradley Fighting Vehicles, and spares. All of the simulators will act as one fighting unit just as if 82 combat vehicles were parked in a motor pool ready to be assigned to a commander conducting an exercise or going into battle.

Once the local networking system is developed, then the research effort will be shifted to long haul networking (LHN) technology. The purpose of LHN is to electronically interplay several local area nets over very long distances without loss of fidelity. SIMNET research will concentrate on three types of long haul networking systems: dedicated digital land lines, wideband satellite, and personal microwave. The major problem which confronts SIMNET in this research is the delay caused from going from a LAN to a LHN and back into a LAN. This delay could potentially present considerable problems in simulation because of the loss of reality.

The SIMNET technology research has been centered around the levels of accuracy needed for various aspects of the simulation, how these relate to specific training objectives, how radio networks can be simulated with its artillery and close air support duties, and how combat support and service support decisions are entered into the network (Herzog, 1986).

The greatest advantage of the SIMNET system is that the soldiers will not fight a computer but other soldiers. The results will be transmitted by the computers via visual and auditory presentations to other players. All of the normal unit procedures such as estimates of the situation, the preparation of operations orders, ground reconnaissance, and reports, serve to inform staff and combatants and to assist the commanders in command and control just as they would in combat of a field exercise (Herzog, 1986).

Until recently, SIMNET research bypassed one of the most crucial areas of the SIMNET system--the area of human performance measurement. However, now ARI's program, "Unit Performance Measurement and Training Program Design for Networked Simulators ", addresses the need to conduct research in the area of human performance measurement in order for SIMNET to become a truly effective training device.

"The performance measurement system for SIMNET will key on relevant ARTEPs and AMTPs to provide a framework for evaluating performance. The specific focus will be the 2A Experimental AMTP being developed for ARI. 2A addresses training objectives to support training and evaluatioon of units at the National Training Center, and uses standard mission scenarios, such as based on a software architecture which collects and analyzes data carried on the SIMNET network, and on inputting observer evaluation data for those tasks which cannot be collected automatically. One of the critical tasks, underway now, is determining the degree to which data can be collected automatically from the SIMNET network, and how much observer/controller input is required." (Vestewig, 1988, p.3).

"The primary goal of the performance measurement system is to help the unit commander by allowing him to review the results of his unit's exercise in SIMNET on hardware available at his home station, and plan additional training accordingly. Ideally, if SIMNET is used before a rotation at NTC, the unit can potentially practice and perfect skills so that the exercises at the NTC are more beneficial. However, the system will also have additional statistical analysis and graphics capability to allow the researcher to perform more sophisticated analysis. Finally the system will have data formats compatible with NTC archives so that analysis can be performed on each data set using the same tools." (Vestewig, 1988, p.3).

2. The Land Battle Test Bed (LBTB)

The Land Battle Test Bed (LBTB) is a wargaming facility which integrates several computer simulations into a single man-in-the-loop, force-on-force simulation. The purpose behind the LBTB is to test new battle concepts with minimum resources. The LBTB is composed of four major components, described below:

* Developmental SIMNET
* JANUS
* An Artificial Intelligence Cell
* ADDCOMPE (Army/DARPA Distributed Communications Processing Experiment) (Shiflett, 1986, p. 50)

The developmental SIMNET system is the major component of the Land Battle Test Bed. "The design features are rapidly reconfigurable simulation modules: semi-automated threat, after action review, automated setup and review capability, and automated user-friendly data reduction (Shiflett, p. 50).

JANUS is a computer assisted man-in-the-loop wargame that is a replacement for the manual "battle" wargame. Only commanders participate in JANUS. The JANUS system is a complementary simulation to SIMNET for examining those issues which exceed the company-sized capacity of the Developmental SIMNET module (Shiflett, 1986).

ADDCOMPE is a joint Army/DARPA distribution communication processing experiment. "Its objective is to provide the maneuver (heavy) proponent with a flexible hardware and software test bed, including supporting personnel, to conduct concept explorations, develop user requirements, and refine issues and criteria for the development of battalion and below command, control, communications and intelligence (C3I) systems (Shiflett, p., 51).

An Artificial Intelligence (AI) cell is also included in the Land Battle Test Bed. Artificial Intelligence is defined for this system as "computer activities which if performed by a human would be considered intelligent" (Shiflett, p. 51). Research in Artificial Intelligence is in the early stage. Eventually, the Artificial Intelligence research will be integrated into the Battlefield Management System (BMS).

In order to demonstrate how each of the components of the LBTB work together, a hypothetical situation will be explained. The issue is extending the area of influence of the battalion commander out to 10km from his battalion center of mass. Several options would be developed by military judgement and the Artificial Intelligence cell. JANUS would be used to determine the desirability of each of these solutions. After the solutions are refined into several variants and organizational structures, they are fed into Developmental SIMNET. Developmental SIMNET would then investigate the system specific parameters using the man-in-the-loop on the close combat heavy battle field. From the work of Developmental SIMNET, the desired system parameters would be described and how to best integrate them into the force.

The Land Battle Test Bed can be used in a number of different applications including:

* A way to study the effects of new technologies.

* The best way to accommodate technological changes.

* How to integrate Artificial Intelligence into armor training and combat development activities.

* A way to shorten the acquisition process (Shiflett, p. 52).

Q. The Company/Team Level Tactical Simulator (COLTSIM)

The Company/Team Level Tactical Simulator (COLTSIM) is a training device being developed to provide a real time, two-sided, computer-assisted battlefield simulation to train company commanders and key subordinates in command and control (C2) of combined arms operations in a simulated combat environment" (Hollenbeck, 1986, p. 53). COLTSIM will become part of the training device strategy to use simulation, substitution, and miniaturization (SSM) to develop and maintain individual and unit proficiency. COLTSIM is being developed to meet the training needs of command leaders and to cope with the great range, speed, and diversity of engagements found in actual com-

bat.

Training methods are divided into two main categories: on-tank and off-tank. On-tank is where the most realistic company/team level leadership training is provided on conventional ranges in areas remote and vast enough to accommodate armor vehicles. Included in on-tank is the Multiple Integrated Laser Engagement System (MILES). Because of the great amount of resources needed for on-tank maneuvers (i.e. equipment, ammunition, personnel, travel expenses, and maintenance) it is usually not possible to have access to such training often enough to become proficient.

The second main category of training (off-tank) consists of manual and computer-assisted simulations. The manual simulators usually have trouble with lengthy game mechanics which greatly reduces the reality of the simulation exercise. Some of the manual systems (Dunn-Kempf and BLOCKBUSTER) cannot appropriately replicate combat stress. The computer-assisted simulators also have some problems. These problems consist of canned scenario and coordination of subordinates. (Hollenbeck, 1986). The COLTSIM system will be able to be used as a stand-alone simulator or integrated into multi-echelon training with other systems being developed under the Army Training 1990 Concept.

In addition, COLTSIM will be able to simulate the following vehicles:

* M113 Armored Personnel Carrier
* M106 Heavy Mortar Carrier
* M60A3, M1, and M1A1 Tanks
* M2 Infantry Fighting Vehicle and the M3 Cavalry Fighting Vehicle.

Also included in the system are controller stations representing higher headquarters, fi[illegible] support, engineer, and the opposing forces. COLTSIM will [illegible] available in both fixed installation and mobile configurations (Hollenbeck, 1986).

The COLTSIM system includes training in troop leading procedures, exercising and practicing ARTEP related tasks, practicing tactics and doctrine, and exercising unit tactical standard operating procedures (SOPs). The average time for warm-up, preoperational checks, and initialization for the COLTSIM system is no more than 30 minutes. However, the most important feature of COLTSIM is that it will provide the training sessions in realtime, with realistic and varied scenarios, and be integrated into other training systems (Hollenbeck, 1986).

R. Simulation in Combined Arms Training (SIMCAT)

The Simulation in Combined Arms Training (SIMCAT) system is a platoon level simulator which incorporates a network of six microcomputers, voice technology, and radio nets. This system provides the opportunity for four students to practice performing tactical scenarios against a free-play opposing force, under the control of an instructor. Line of fire and movement rate variables are computer controlled, and the students move and fire graphic representations of their tanks using normal voice commands. The SIMCAT system provide a bridge between the classroom training and the field exercises (Burnside, 1986).

S. National Training Center (NTC)

No discussion of major Army training strategies and training settings would be complete without a description of the U.S. Army's National Training Center (NTC) at Fort Irwin, California. NTC is the most sophisticated, battalion-scale, combined arms training system in the world.

NTC provides an instrumented battlefield covering 636,000 acres. Battalions from locations across the country are brought to the center in 20-day rotations to spend 10 days in force-on-force engagements using MILES and 4 days in live-fire exercises. The battalion faces opposing forces, outnumbered by a ratio of 3:1. NTC has a full-time professional group of referees and members of the Red Force command numbering 530. Missions vary to include all critical missions, such as movement to contact, hasty attack, and deliberate attack.

The instrumentation system includes realtime battlefield monitoring and control for two simultaneous, combined-arms force-on-force engagement exercises. The battalion live-fire exercises provide capability against 500 tank targets, computer-controlled threat array, with automated scoring. Monitoring and recording is provided of real-time position location, tactical communications, hit/kill/near miss events, and weapons system pairing for 500 players. The force-on-force engagement scenarios, performed day or night, use Multiple Integrated Laser Engagement System (MILES). The monitoring and recording of command, control, and tactical VHF communications takes place on 90 selectable channels.

The center also utilizes real-time video recording and display from eight mobile video units, two backup video/Tactical Operations Center (TOC) units, and two long-range video surveillance systems.

The instrumentation system provides real-time graphics displays of specific battlefield activity and statistical data, overlaid on digitized terrain maps. The NTC Operations Center is equipped with two exercise control rooms, briefing theater, and digital/audio/video monitoring, recording, and editing facilities. Systems analysts stationed at the Operations Center integrate input from the control room displays with input from the observer/controllers stationed throughout the battlefield. Two mobile after-action review vans are used to provide immediate training feedback to the unit after each exercise.

Future growth projections for NTC include system expansion to accommodate 1,000 players, air-to-ground engagement simulation/air defense training, and expanded live fire training with 1,000 automated tank targets.

T. Summary

This review of current Army military team training approaches is illustrative of the emergence of high technology into the military training arena. Comparable Navy, Marine, and Air Force approaches to team training are presented in Appendix A. There are virtually no team training systems that have been described that do not incorporate simulation techniques, and/or computerized generation or retrieval of training data. The use of simulators has, in fact, become the reality of the training function due to excessive costs and logistical and practical problems of real-life demonstrations of combat performance. Current computerized team training approaches have definite advantages, particularly as they provide important automated performance measurement capabilities that vastly reduce the pressures for evaluators to attend to all aspects of trainee performance. At the same time, as shall be seen in the next chapter, there are associated problems with computerized training techniques, one being that valuable and integral team performances involving group interactions and reactions to noncomputerized, emerging stimuli may be lost if not attended to by trained observers.

Chapter IV. Current Army Evaluation Systems and Other Emerging Measurement Systems

The concepts of measurement and measuring unit performance have been described previously. It is the purpose of this chapter to elaborate on the topic of performance evaluation as it relates to the systems of team training which were described in the previous chapter.

According to Spencer, Klemp, and Cullen (1977), information on Army evaluation is readily available and plentiful. However, the evaluations are most often vague and general, with actual measures, correlations, and other statistics seldom being stated. Spencer et al. (1977) list the following types of Army performance effectiveness criteria: Inspection Scores; Mission Accomplishment Results; Efficiency Measures; and Personnel Development Indices. In this chapter we will primarily focus on observational evaluations, exemplified by ARTEP scores for light infantry battalions (ARTEP, 1985). An overview of other types of Army performance effectiveness criteria listed by Spencer (1977) is presented in Appendix B. In addition, the emerging technologies associated with SIMNET and COLTSIM are reviewed to point out the need to include adequate performance measurement as part of these systems. Finally, several emerging performance measurement techniques will be reviewed.

A. ARTEP Scores

ARTEP is the most commonly cited performance rating in the Army consisting of: (1) mission/task oriented training and evaluation, (2) concurrent, multi-echelon training and evaluation, (3) training to correct deficiencies, and (4) decentralized training and evaluation. According to Havron et al. (Vol. II, 1978), for ARTEPs to be most effective, they should be as decentralized as possible. ARTEPs may be used to evaluate single battalions during training, or may assume a more integrated multi-unit approach to training.

ARTEP plans emphasize that the command group and staff should be fully trained beginning the simulation exercise. However, in some of the ARTEP plans (e.g. Light Infantry Battalion and Brigade Operations and Battalion ARTEP, 1985) no attempt is made to establish levels of proficiency.

Commanders are encouraged to use the ARTEP Mission Training Plan (MTP) to conduct frequent internal evaluations of the

command group/staff's performance. These evaluations should provide the commander with the information needed to develop or modify the training program. Evaluations are guided by the following principles:

(1) Development of the evaluation plan. The commander is responsible for the development of an evaluation plan for internal evaluations. The senior evaluator has the same responsibility for external evaluations.

(2) Training of evaluators. The evaluator training program is developed with supervision from the senior evaluator. The program stresses specific evaluation requirements to develop a thorough understanding of the tasks that are to be evaluated and the training/evaluation standards which must be met. If the evaluations are attained from a particular training device, then the evaluators are thoroughly trained on the mechanics of that particular device.

(3) Preparation of test documents. Either the commander or the senior evaluator is responsible for preparing the overall operation scenario based on specific terrain where enemy threats would exist. The standards outlined in the Training and Evaluation Outlines (T&EOs) guide the development of the scenario. The commander or the senior evaluator is also responsible for the preparation of all warning and operations orders required in the T&EOs, and also a control plan which includes evaluator assignments and responsibilities, evaluator communication, evaluator reporting, required reports, and instructions concerning evaluator training.

The T&EOs contain the tasks which are used in battle simulation for evaluation. An important part of each simulation is the evaluation guide which can be used to judge the ability of the command group and staff in accomplishing specific tasks.

According to ARTEP FC7-13 for Light Infantry Battalions (1985), the development of the evaluation strives for a valid evaluation of a unit at an acceptable cost in resources. The plan must fit the situation and the needs of each unit. However, the guidelines state that there is no "best" evaluation plan, only general approaches which should be considered.

In order to meet the objectives of the evaluation, the evaluators:

* Brief the unit to be evaluated on the evaluating concept.

* Conduct and control the evaluation according to the conditions in the ARTEP.

* Fit the sequence of events to the terrain and to the training environment.

* Identify the critical points during the evaluation where command group task accomplishment can be objectively assessed, i.e., where to measure "product" rather than "process".

* Rate task performance objectively to ensure consistent application of performance standards.

* Rate actual command group performance according to the standards listed in the T&EO and current field circulars.

Evaluator training is important and, in order to qualify evaluators to conduct ARTEP evaluations, evaluators need to:

* Review the ARTEP to clarify its contents and establish evaluator objectives.

* Develop scenarios and an evaluation plan.

* Conduct an oral review of the scenarios and evaluation plan to develop courses of action and interpretations of the evaluation standards.

* Conduct a map exercise of the actual test.

ARTEP Mission Training Plans (MTP's) are training publications that serve as guides to the development of a training program. However, such a program will only be effective if it can overcome training deficiencies. These deficiencies are discovered in a number of ways; usually by observation and other more structured evaluation methods. The ARTEP Mission Training Plan provides guidance for such evaluations.

The senior evaluator is responsible for development of the evaluation plan and scenario. The evaluator chooses the appropriate tasks and organizes them into a logical sequence within the tactical scenario which is consistent with the contingency mission. In the evaluation of the participants, there is no mathematical formula which is used to arrive at a

pass-fail grade. If an overall rating is given, then it consists of a satisfactory/unsatisfactory rating made by the senior evaluator based on observations of all the separate performances and analyzing the results of those observations with professional judgement. Ratings are consistent with training management terminology (i.e., unit performance satisfactory, unit performance unsatisfactory, unit is trained and ready, unit needs practice, unit status is unknown and unit needs substantial training) rather than numerical ratings. Evaluation using the ARTEP plan does not produce a readiness rating. Performance in any evaluation is considered only one factor used by the commander to assign an overall readiness rating. If any deficiencies are identified during evaluation, then the battalion commander should design a training program to overcome them.

Because so much responsibility is placed on evaluators,it is important to select only experienced and qualified evaluators because they must apply sound professional judgement when analyzing the performance of the command group/staff. The senior evaluator, ideally, would have past success in commanding a similar unit. The senior evaluator is responsible for the selection and training of assistant evaluators and ensuring that the battle controllers are properly trained. Finally, the senior evaluator must be very familiar with the evaluation standards.

The battalion ARTEP is a performance-oriented test administered to the battalion by its higher headquarters (Brigades or Division). During the 72-hour evaluation period the tactical proficiency of the battalion's headquarters as well as the headquarters and headquarters company is assessed. The purpose of the evaluation is to provide the commanders a means to determine the level of command and control proficiency of battalion headquarters and separate elements of the HQs and HQs Company and subordinate companies. The senior tester and his assistants observe and record the performance of leadership and teamwork of the commander and staff, using selected T&EOs. The analysis of the test data provides feedback on strengths and weaknesses of the battalion headquarters and separate elements and form a basis for future training and resource allocation. The ARTEP plans provide general guidelines and procedures for planning and scheduling the battalion test. However, the general guidelines must be adapted to local conditions and command prerogatives.

ARTEP guidelines state that at least ten of the twenty battalion missions should be used in the test in order to provide an adequate test bed to evaluate the battalion's ability to perform. Consideration in selecting the missions should be

given to the degree of difficulty from both the execution and evaluation standpoint.

The evaluators guide the evaluation through the opposing force (OPFOR) actions that are contained in the ARTEP test scenario. The strength ratios for tasks contained in the test scenario for OPFOR vs. player unit is controlled by the senior evaluator. Evaluators observe player and OPFOR activities and determine whether the tasks are performed to standards contained in the T&EO's. OPFOR may come from the unit or from external sources.

The evaluators are given a packet that provides information about the selected missions. The packet contains: Training and Evaluation Outlines (T&EOs), T&EO and mission accomplishment summary, overall test summary, unit data summary, environmental data report, casualty report, and Operations Orders (OPORD).

An administrative site is required on the field test site in order to stockpile training ammunition and other needed equipment. This site is also good for debriefing evaluators and unit members, and as a central location for test data to be collected. Usually the brigade command post can be used to perform this function.

The senior evaluator controls the exercise through the command post and through the use of a radio net. The command post (Test Control Headquarters) should be organized as a stationary or mobile company command post (CP) to exercise control and portray realistic time-distance factors between the brigade CP and the battalion.

The senior evaluator terminates a module when the battalion has completed all the missions/tasks in the module or have suffered enough casualties so that the missions/tasks cannot be completed (this is determined by MILES information). A record is made of the reasons for the termination. After the termination of one exercise and before the next mission is undertaken the following actions must be performed by all evaluators:

* All MILES equipment must be promptly inspected, kill codes recorded, and then reset. Any damaged or inoperative MILES equipment must be replaced.

* The evaluators must promptly resolve all casualty data to determine the time, place, number, and cause of casualties. This information, along with the GO/NO GO/NOT EVALUATED information from the

T&EOs, must be reported to the recorder in the test control headquarters.

* The evaluators must debrief the commander and platoon leaders to clear up any questions. The senior evaluator will then direct the battalion to continue its mission once brigade fragmentary orders (FRAGO) or operational orders (OPORD) for the next module has been issued.

The evaluator's duties and responsibilities during conduct of the test are to:

* Check GO/NO GO blocks of T&EOs to record whether or not standards were met.

* Collect data required in the reports supplied in the tester packets.

* Ensure that all MILES equipment is functional.

* Report major kills (groups of personnel).

* Report major weapons firing.

* Enforce rules of engagement.

* Observe critical tactical events.

* Record routes of travel and unit locations.

* Enforce safety.

The evaluating team controls the Battalion Test in two ways: first, through the control measures established in of the OPORD or FRAGO; and second, through the brigade commander (simulated by the senior evaluator) on the brigade command net. The main purpose of the evaluating team is to observe. The team members do not aid the battalion commander in any way. Evaluators must remain neutral throughout the test.

After the test has been terminated, the battalion is moved into an assembly area and the following actions should be performed before moving back into garrison:

* The senior evaluator must debrief subordinate evaluators and compile all data (evaluator packets).

* The senior evaluator must complete the test summary report to reflect overall unit performance, using

the evaluator scoring system.

* All complete evaluator packets must be turned into headquarters for recording and analysis.

* The senior evaluator must conduct an After-Action Review of his element's performance.

After the battalion has completed the test, the senior evaluator collects all evaluator packets used by the evaluation team. The senior evaluator completes the overall test report by using his own observations as well as those of the evaluation teams. The report consists a measure of overall battalion proficiency which includes the following:

(1) Training and Evaluation Outline (T&EO). The use of the T&EO as an evaluation tool is crucial in assessing unit performance. The standards of the T&EO provide information in determining the strengths and weaknesses as related to specific categories.

(2) Mission Accomplishment Summary. The mission accomplishment summary (A-B-C-D scale) is used as an additional indicator of unit proficiency and assesses accomplishment of each mission. These standards are based on an assessment of MILES casualties and equipment (e.g., for platoon criteria, 20 killed in action (KIA) or wounded in action (WIA) constitute "light casualties").

(3) Unit Data Report. This report presents demographic information which reflect on a unit's performance (e.g., new leaders, low strength, etc.).

(4) Environmental Data Report. This report presents weather information so that a comparison of missions conducted under different environmental conditions can be made.

(5) Casualty Report. This report presents information which reflects on a unit's degree of success during engagements with OPFOR.

(6) OPORD. The OPORD is used by the senior evaluator to begin the test scenario.

The T&EOs and the Mission Accomplishment Summary (MAS) are the two sources which support the scoring system. The T&EOs demonstrate the battalion's ability to perform the tasks to the established standards. For example, an overall GO for the

T&EO for each task may be 10 points and NO GO be 0 points (taken from performance summary of T&EOs and based on the evaluator's judgement). The mission accomplishment summaries show the battalion's ability to sustain performance. For example, A = light casualties, B = moderate casualties, C = heavy casualties, D = mission not accomplished. In other words, the unit may be able to perform the tasks to standards, but may sustain too many casualties. This is what the MAS is designed to measure. The specific procedures that the senior evaluator uses in compiling the final score are as follows:

(1) The senior evaluator compiles all T&EOs and mission accomplishment summaries related to each mission and calculates the unit's score using a point system.

(2) The senior evaluator adds up the points for T&EOs and determines the MAS score. This score represents the alpha numeric battalion score for the particular mission. The battalion will receive a separate score for each mission executed during the test.

Although there are many rich and varied opportunities for analysis of ARTEP data to assess potential predictors of good or poor performance, unfortunately there are very few reports, other than those described in the next section, indicating that ARTEP scores are used for research purposes.

B. Combined Arms Tactical Training Simulator (CATTS)

In one of the few tests of newly emerging systems using ARTEP scores, Thomas, Barber, and Kaplan (1984) obtained ratings of battalion command group (BCG) performance and of the exercise difficulty and realism during Combined Arms Tactical Training Simulator (CATTS) exercises both from self estimates by the command group and from assessments by the player-controllers and controllers. Raters were asked to assess each of the subtasks relative to a standard and assign a number to each subtask that reflected how many times greater or lesser it was than the standard. The standard, defined as the minimum acceptable performance in a tactical environment, was assigned the value of 100.

Performance estimates were obtained after each of the 20 CATTS training exercises. In addition, after each exercise, a rating estimation was obtained concerning the difficulty and realism of that exercise. A relatively high inter-item correlation was attained which suggests that the raters had difficulty discriminating among the ARTEP items intended to assess battalion command group performance and exercise difficulty. However, it appears that the players and player-controllers

were better able to make these discriminations than the self estimates by the command group.

In addition to analyzing the relationship between items, the amount of agreement among raters in the performance of command groups was also examined. First, inter-rater agreement was calculated by a comparison of ratings on the 27 ARTEP performance items for each exercise day both among and between rater groups. The correlations were highly variable (ranging from r = -.97 to r = +.94) and on the average, low (median r = .10). There were no consistent patterns among any of the raters. These findings are consistent with the earlier results and suggest that the raters could not make the fine item discriminations required by the ARTEP instruments.

While there was some general agreement within a subset of controllers, there was not a consistent agreement among all controllers as to battalion command group performance. The authors suggested that the general lack of inter-rater agreement among controllers (and by extrapolation players, and player-controllers) might be due to differing interpretations of the ARTEP standards. The authors concluded, "These findings coupled with the lack of item discriminability highlight the need for modification in the performance rating instruments" (Thomas, Barber & Kaplan, 1984, p. 17).

Modified versions of two objective indices obtained from combat development studies, Relative Exchange Ratio (RER) and Surviving Maneuver Force Ratio Differential (SMFRD), were used as measures of battlefield performance. These two measures of performance were selected because they were found to be significantly correlated with controller ratings of overall command group performance during previous CATTS exercises (Thomas, 1984).

All of the realism ratings made by the observers (controllers, players, and player-controllers) were higher than the standard for minimum realism perceived by raters to be effective in a training exercise. This indicates the fidelity of the CATTS system in all the areas surveyed, as assessed by these ratings, was at least adequate and sometimes very good.

However, the controller ratings of performance were highly correlated with ratings of realism (typically in the mid .80's), which suggests a possible bias in controller ratings. Player-controller correlations between the ratings of realism were also significant, but typically not as great (usually about r= .60). The correlations for player ratings were even lower, typically in the mid .40's. These results suggested to the author that performance ratings were differentially

reliable across raters.

In addition, manipulation of the system characteristics had a strong impact on simulation outcomes of the CATTS system. These variables need to be set so that different combinations of combat ratio, mission weather, communications, and terrain can result in exercises with approximately the same level of battle difficulty. Only then can performance change as a result of CATTS exercises be attributed to factors other than exercise difficulty. Perception of battalion command group (BCG) performance and exercise difficulty were not affected by the system and scenario manipulations. The experimental design did not allow for direct comparison of these measures, as raters and experimental conditions were confounded. The controllers did observe all of the exercises, but their ratings indicated an inability to discriminate among and between items intended to measure BCG performance and exercise difficulty.

The researchers conclude: "It appears that further research is required to develop subjective ratings of BCG performance that are both reliable and valid. Steps should be taken to develop new, more objective measures of C^2 performance, such as a series of naturally occurring battlefield events (probes), that could be used to determine how well BCGs exchange this critical information in a timely fashion during CATTS exercises. The use of performance appraisals by non-involved, outside tactical operations center observers such as tactical operations center (TOC) monitors to assess C^2 behavior may produce more reliable and valid measures" (Thomas, Barber & Kaplan, p. 31).

Another recent study, although not designed to specifically study the reliability or validity of ARTEP scores, was conducted to determine if the ARTBASS test bed, CATTS, provides effective training in the collective and individual critical tasks of battalion command groups as specified in ARTEP 71-2 (Miller & Noble, 1983). The study sample consisted of twelve (12) regular army battalion command groups and concentrated on the five principal staff officers of each group: the battalion commander (Bn Cdr), the personnel and administrative officer (S1), the intelligence officer (S2), the operations officer (S3), and the logistics officer (S4).

Data were collected prior to training, during training, immediately following training, and 30 to 45 days after training. The data included group performance ratings, player perceptions of training effectiveness and realism, self-estimates of performance, a written proficiency test, measures of intragroup relationships, and background information on the study participants.

Analysis of the group performance data was restricted to ratings from four systems, by controllers who were highly trained, qualified professionals. A measure of concordance indicated the controllers had applied essentially the same standards in ranking the performance of training groups (W = .55). Based on system controller ratings, training groups showed significant performance improvement across training exercise in most areas. Self-estimates of performance significantly improved for all staff members, except the S4 who, according to player comments, was not sufficiently exercised by the simulation. All collective tasks and a majority of the position-specific individual tasks were rated as well-trained. However, player ratings indicated that certain individual ARTEP tasks were not well tested by the simulation. Most of these tasks were related to security measures, and the authors speculated that these low ratings would probably not occur with the ARTBASS because training would be conducted in field locations using the units equipment and Standard Operating Procedures (SOPs), as opposed to the fixed garrison location of the CATTS. The CATTS exercises also promoted team building behavior. Worth noting is that these training benefits (e.g., "team building") may indicate a need for such simulation training in order to promote this team-building behavior.

Miller and Noble (1983) point to several constraints imposed on the study which must be weighed in accepting the conclusions. There were no comparative control group measures and no measures of training transfer. In addition, the findings are based on the subjective estimates of the study participants and the system controllers. Given the fact that performance improvement was desired and expected, the possibility of some institutional bias towards improved ratings cannot be ignored.

Within these limitations, the authors concluded that ARTBASS, as represented by the CATTS test bed, could be a valuable addition to the Army's inventory of battle simulations. Although several training issues remain unaddressed, specifically those of cost, training transfer, and comparisons with control groups or alternative training systems, they also recommended that an ARTBASS post-fielding training effectiveness analysis be conducted following initial operational capability.

C. After-Action Reviews (AAR)

After-action review (AAR) is an important part of the effectiveness of many arms training techniques, including ARTEP.

However, at the NTC, which utilizes the most advanced form of AAR available, including multi-media take-home packages con-

taining complete summaries, audio/video graphics presentation, and written documentation of what was done well and what to improve on, ARTEP standards are not applied. AARs at NTC concentrate on doctrinal review to reinforce the training aspect, as opposed to evaluation aspects.

The controller's responsibility is leading the review, facilitating the exchange of information, and promoting discussion by posing appropriate questions. Without his overall knowledge of the exercise, the review has a tendency to be brief, and much information which should be exchanged by the players can be overlooked. In order for the training to be effective, the review must be considered just as important as the actual conduct of the exercise.

The reliability and validity of after-action reviews has not been given much attention. The "effectiveness" is determined by the controller. The exercise controller is cautioned not to use mission accomplishment as the only criterion for evaluating tactical decisions. The results in a given situation, therefore, can be attributed to either good execution by one force, poor execution by the other, or a combination of the two. There is no indication of a reliable method for determining good, average, or bad execution.

D. Simulation Networking (SIMNET)

The preceding evaluation systems represent relatively traditional observational procedures. Newer evaluations systems rely heavily on automated equipment. For example, SIMNET will use the most modern compact computers in the evaluation of the system. These microcomputers will model and monitor the performance envelopes, maintenance conditions, combat damage, and consumable status of each vehicle, as well as compute fly-out of projectiles. Computers will also be responsible for the location of all vehicles in the simulation as well as firing events and weapons effects. The computer's basic underlying role is to translate human interactions into visual and auditory presentations of sufficient fidelity to cue the appropriate behaviors on the part of the combatants (Herzog, 1986).

The basic SIMNET system is designed to be flexible due to the computerized network data base. In addition, the SIMNET system can be set up from a tank-on-tank experience to a platoon-on-platoon, which would involve all members of the network.

The SIMNET system does, however, contain some shortcomings. Three of the shortcomings that are important to this re-

port are: after-action review, feedback, and human performance evaluation. After-action review is necessary so that the training manager can observe a record of the participants activities and ensure that proper action was taken. The key to the SIMNET system is to increase practice on those critical individual, crew and collective skills and leader tasks that will lead to success on the battlefield (Herzog, 1986). However, the SIMNET system leaves out two of the most important aspects of accomplishing better battlefield performance: appropriate feedback to the participants and appropriate human performance evaluation.

These deficiencies are receiving preliminary attention in the Land Battle Test Bed (LBTB) facility. The developmental SIMNET system is the major component of LBTB which is designed to feature rapidly reconfigurable simulation modules: semiautomatic threats, after-action review, automated setup and review capabilities, and automated data reduction. In the SIMNET system, the opposing players are each others threat. However, in order to increase the reality of the simulation, an automated opposing force (OPFOR), perhaps controlled by one or two trained players, is absolutely essential. The after action review makes reference to the ability to review the processes involved in the conduct of the battle. "Our experience with other wargames, and in a field experiment or FTX, indicates that the output data in the form of traditional force effectiveness measures is not sufficient to explain the processes which produced them (emphasis added). A means of replaying the battle from a host of different viewpoints is essential" (Shiflett, p.50).

The two features of an automated setup/review capability and automated user friendly data reduction are for analysis purposes. In a highly complex system, such as LBTB, however, the developers must beware of the potential problem of gathering too much data to make the system usable. Enough data must be provided to make the system a worthwhile training simulator, but too much data can make analysis unmanageable. In addition, although the system provides automated review and data reduction, it still relies on the judgement and expertise of the trainer to provide appropriate feedback and evaluation to trainees.

E. Company/Team Level Tactical Simulator (COLTSIM)

The Company/Team Level Tactical Simulator (COLTSIM) is still in the development stages. However, in order to train and evaluate command and control skills, COLTSIM requires certain characteristics. The system should be capable of accepting and responding to operator inputs, of processing large vol-

umes of data in realtime, and have intercom and external radio transmission. The commander and controller should be able to receive visual and audio outputs for monitoring. One of the most crucial characteristics that the system should provide is a hardcopy record of actions as a means of evaluating and critiquing performance through after action reviews. This will enable the commander to learn how to evaluate the combat performance of his unit while performing his own duties in the heat of battle (Hollenbeck, 1986). At this point, however, these evaluation techniques are largely incomplete.

F. JESS, JANUS/JANUS (T), ARTBASS, ASSAULT, and COMBAT-SIM

The purpose of the Battle Simulation Software Survey (1986) undertaken by PM TRADE, was to document the results of a market survey of battle simulation software that might be used to satisfy the requirements for Division Battle Simulation (DBS), Brigade/Battalion Battle Simulation (BBS) and Combat Service Support Training Simulation System (CSSTSS). It identifies catalog information on candidate battle simulations, compares them to the requirements and provides a preliminary comparative analysis. According to their BBS comparison chart, all the evaluated systems (JESS, JANUS, ARTBASS, ASSAULT, and COMBAT-SIM) fulfilled the requirements for checkpoint, restart and replay capabilities, provide video capability at each work station, and provide an in-depth after-action review, with threat and friendly forces modeled separately, except for the ASSAULT system which only partially meets requirements for checkpoint, restart and replay capability and does not provide video capability at each work station.

G. The National Training Center (NTC)

As described previously, the NTC uses objective, realtime, training feedback data in the form of graphics displays, audio, and video presentations produced by the Range Data Management Subsystem (RDMS) and Range Monitoring and Control Subsystem (RMCS).

NTC personnel provide immediate after-action reviews in the field after each exercise, usually taking up to 2 hours. The feedback is an integration of observations, battlefield monitoring from observer/controllers (O/Cs) and mobile video units who send communications from the field to the NTC Operations Center, and output at the Center where analysts see the battles through computer and video workstations linked with O/Cs in the field. Analysts also have access to all communications nets and choose which to record. All input is recorded and time-tagged for future review in AAR vans for post exercise critique. The AAR is also recorded.

There are several weaknesses which have been identified in the current system; automatic data collection/recording is limited to RDMS data, observer/controller (O/C) input to the NTC data base is negligible, and exercise control is limited to FM voice radio. What is needed for a new system is to establish a two-way digital data link between the OPS and the field O/C. The device should be compact for vehicle mounting and limited man-portability, provide direct input from the field O/C into the NTC database, and provide OPS center graphics to the field O/C. This should help the O/C provide formatted or structured feedback to troops at the lowest level (Industry Day at the NTC, 1987).

H. Summary of Current Army Evaluation Systems

The Army uses a large variety of evaluation techniques, the primary one for unit performance measurement being ARTEP. There are still problems with the ARTEP system, however. The system requires extensive ARTEP experience for evaluators to interpret performance and evaluate it accurately. There is little reliability to the system because evaluators fail to discriminate among ARTEP items, and standards of performance are vague. The NTC provides good after-action review, but does not provide feedback in relation to ARTEP standards and does not provide reliability or validity evidence for training effectiveness. Newly emerging systems, such as SIMNET and COLTSIM, offer automated data reduction and review capabilities, but do not incorporate sufficient evaluation against standards which provide trainees with meaningful feedback. Here again, the quality of evaluation feedback relies on the expertise of the trainer.

Many of the evaluation and measuring techniques described above are not mutually exclusive to the Army. Nor are measurement problems. Because many measurement systems in other services overlap those of the Army, a brief overview of evaluation systems used in the other services is provided in Appendix C.

I. Other Emerging Measurement Systems

1. Joint-Service Performance Measurement/Enlistment Standards Project (Project A)

The military is currently investigating ways to improve performance measurement at the individual level. In July 1980, a Joint-Service effort was launched to investigate the feasibility of measuring on-the-job performance and using these measures to establish military enlistment standards. In 1983, a

Joint-Service research program was established: 1) to develop prototype methodologies for the measurement of job performance, and 2) if feasible, to link job performance measures with enlistment standards. The Joint-Service Job Performance Measurement Group is the primary source of review and coordination of Service job performance measurement research programs (Harris, 1986).

The overall strategy is for each Service to demonstrate its ability to efficiently collect valid, accurate, and reliable hands-on job performance measures. These measures will then be used as benchmarks to evaluate surrogate (less expensive, easier to administer tests and/or existing performance information) indices of performance as substitutes for the more expensive, labor intensive, hands-on performance measures. (The use of surrogate measures will be discussed in Chapter VI Research Implications.) Each Service is developing key components of the overall Joint-Service program within a common methodological framework.

The Army is lead Service for:

* Hands-on performance tests for one cross-Service Military Occupational Specialty (MOS) (Military Police) as well as hands-on tests for Army-specific MOSs.

* MOS-specific job knowledge tests.

* Army-wide performance ratings.

MOS specific job performance measures include hands-on performance measures, measures of training success, and job knowledge tests. Army-wide performance measures using behaviorally anchored job performance rating scales are also being developed.

The Air Force is the lead Service for:

* Hands-on performance tests for the Jet Engine Mechanic and Air Traffic Controller, Air Force Specialties (AFS) as well as for Air Force-specific AFSs.

* Walk-through testing development and demonstration.

* Job experience rating development and demonstration.

* Cross-Service use of performance measurement strategies.

The objective walk-through testing is designed to expand the range of job tasks assessed to include tasks which do not lend themselves to hands-on testing because of cost, time, and/or safety considerations. Walk-through testing is a task-level job performance measurement system which combines task performance, walk-through and interview procedures to provide a high fidelity measure of individual technical competence. The walk-through testing methodology is being evaluated both as a supplement to hands-on data collection and as a more cost-effective substitute (i.e., surrogate).

A wide range of rating forms are being developed as alternative job performance measures in addition to the interview testing methodology. These include peer, supervisor, and self performance ratings at four different levels of measurement specificity -- tasks, dimension, global, and Air Force-wide -- as well as ratings of job experience.

The Navy is the lead Service for:

* Hands-on performance tests for one cross-Service rating (Electronic's Technician) and for Navy-specific ratings.

* Simulation performance test development and demonstration.

* Symbolic simulation substitute test development and demonstration.

* Comparison of various Service-developed measurement strategies within a single career field.

Hands-on performance tests will use actual equipment or parts of equipment in assessing technical proficiency and may involve whole or part task sequences. One type of substitute for the hands-on test will use either "low" fidelity computer-based technologies, such as videodisc systems, or paper-and-pencil simulation techniques that rely on pictures and illustrations of actual equipment. The second type of substitute will use behaviorally anchored rating scales that specify work behaviors representing different levels of proficiency.

The Marine Corps is the lead Service for:

* Hands-on performance tests for one cross-Service MOS (Automotive Mechanic) plus Marine Corps-specific MOSs.

* Identifying the total range of information all Services must include in their data acquisition efforts.

The Marine Corps will develop two types of tests to measure job performance: hands-on performance tests and surrogate written tests. They will also attempt to adapt ratings of performance that have been developed and evaluated by other Services.

According to LT Col D. A. Harris, the DOD Project Manager (1986), demonstration of prototype performance measures and initial attempts to link those measures with enlistment standards are expected in 1987. These developing performance measurement systems, although targeted toward individual assessment, will certainly influence developing collective performance measurement systems. For example, will Project A be successful in developing reliable hands-on measures and of potential surrogate measures? If so, will these measurement methods relate to collective training?

2. Measurement of Team Behavior in a Navy Training Environment (TEAM)

Recently, research has been reported from a project to measure team evolution and maturation (TEAM) as team members gain experience and knowledge about tasks, each other, and external environmental demands within the context of an operational training scenario (Morgan, Glickman, Woodard, Blaiwes, & Salas, 1986). In addition to using several observational and interview techniques, three data collection instruments were developed specifically for the research including the collection of critical team behavior reports from instructors, self reports of changing perceptions from team members, and a demographics form. The critical incidents form will be described here in detail because it has been used intensively and successfully to determine training needs, curriculum design, and performance requirements in the Navy (e.g. Glickman & Vallance, 1958) and elsewhere.

In this study, a critical incident approach (Flanagan, 1954) was used to develop a critical Team Behaviors Form to be used by instructors as a means of identifying specific effective and ineffective behavior of team members. The first step in developing critical incidents was to conduct semi-structured interviews with instructors to extract critical behaviors that seemed to fit the dimensions of the TEAM model. The critical incidents were then content analyzed and categorized into seven dimensions: communication, adaptability, cooperation, accep-

tance of suggestions or criticism, giving suggestions or criticism, team spirit and morale, and coordination. Critical team behaviors were then dichotomized as effective or ineffective so that each page of the Critical Team Behaviors Form contained either the effective or ineffective behaviors of a given dimension. Instructors were asked to place an X in the box under the position of each member involved in an observed critical behavior. Responses to the forms were examined to compare characteristics of teams that were regarded as being effective with those that were judged to be less effective and to identify those behavioral dimensions that were most sensitive to the differences between good teams and poor teams over time. Results showed that it was indeed possible to discriminate a "good" team from a "bad" team using the TEAM methodologies. For example, the frequencies of effective and ineffective behaviors obtained from the critical Team Behaviors Form provided meaningful comparisons of teams, sessions, and dimensions of behaviors. Use of the critical incidents methodology, therefore, offers good promise for the evaluation of performance in collective military units.

3. The Headquarters Effectiveness Assessment Tool (HEAT)

Another emerging system has been developed for measuring the effectiveness of joint/combined training in contrast to measuring the development of team behavior as exemplified by TEAM methodologies. The Headquarters Effectiveness Assessment Tool (HEAT) was developed with the recognition that effective and efficient unit training requires explicit, quantitative measurement, and (b) measurement techniques (e.g., ARTEP) exist for battalion and below, but tend to be binary rather than continuous, quantitative performance-based measurement.

HEAT was developed in 1982 by Defense Systems, Inc. because of the need that existed for achieving reliable and valid measures of performance of command and control (C2) systems. Originally it was part of a project that also developed guidelines for the design of such systems at the headquarters level, and HEAT was intended as a tool that would help qualify the performance of a headquarters. However, the HEAT system was discovered as being suitable for application to headquarters and command nodes at other levels.

HEAT has been applied in a number of different settings including field exercises involving joint forces (including Army Division size elements), naval battle group and fleet exercises, by the Military Airlift Command in a variety of laboratory experiments in C^2. Even though the technique must be modified for applications with different settings, it is the underlying methodology that consistently provides quantitative,

and reproducible assessments of the quality of the C^2 processes observed.

Since HEAT measures the performance of separate components and processes as well as overall headquarters performance, it can be used to measure the impact of new equipment and the possibility of creating new procedures, training methods, or doctrinal insights.

The Headquarters Effectiveness Assessment Tool (HEAT) combines measures of effectiveness impact on military situations along with diagnostic measures of the quality of the processes that were used within the headquarters to attain the results. The HEAT system measures also include measures of the time required for key processes and the linkages of those measures to the relevant components of the environment. It can also be applied to any set or subset of headquarters functions, with measures of individual tasks being weighted according to their relative importance.

HEAT is designed to assess the quality of the processes used by the headquarters (monitoring, understanding, planning, and directing) by monitoring a headquarters' interactions with the environment it seeks to control and its ability to convey and integrate the appropriate information quickly and correctly. The headquarters is responsible for the development of "plans" for the subordinate, which are defined as a set of missions, assets, and schedules developed for subordinate commands. The effectiveness assessment is based on the ability of the headquarters to develop and implement such plans while adjusting them for information and assets available. Effectiveness would be hindered when too much is attempted by too few forces, too quickly or in the absence of adequate information. Knowledge of these deficiencies is attained from the action of the headquarters' own decisions to alter the plans. Therefore, the actions of the headquarters, not the observers, provide the information that is necessary for the effectiveness evaluation.

The cycle of the headquarters consists of six process steps. HEAT includes measures of the quality of these processes as a diagnostic tool, described below. However, a high quality of processes does not necessarily produce a high headquarters performance. The HEAT system does allow, when headquarters performance is particularly weak or strong, an analysis of the processes that were performed well or the processes that need to be improved.

The variables that the HEAT system measures in relation to headquarters effectiveness ratings are the ability to monitor,

to understand, to choose alternative actions, to predict, to make sound decisions, to give direction, and to interact with the environment.

In order to properly apply the HEAT system the user must define:

* Areas of study, i.e. , the specific headquarters functions and process steps to be studied;

* Events to be covered, i.e., a specific exercise operation or period of time; and

* The specific HEAT measures to be applied.

In addition to defining the above items, the user of HEAT must also provide specific standards for many of the activities to be measured and assign weights to the individual measures reflecting their relative importance in the application.

The HEAT observers must be trained to understand the basic concept of HEAT so that all of the significant information will be recorded. The observers must be trained in a manner that would allow them to recognize the headquarters cycles and identify the steps that were used in the cycle. Therefore, the first step in post-exercise data collection would be the establishment of a HEAT oriented chronology of exercise events, in which the cycles are identified.

The scoring of the HEAT system produces two types of scores. First, a raw score is derived which corresponds to the definition of the measure, and secondly, a normalized score is derived on a scale of 0 to 1 which is the raw score compared to the performance standards supplied by the user.

The human observer plays an important role in the application of the HEAT system. The data collected through observation are used in addition to the exercise-generated data or the actual physical records obtained. The observers usually gather data relating to the headquarters' understanding of the current situation, courses of action being considered, decisions (selection of a course of action), and predictions of a decision's outcome or the implications of alternate courses of action. This type of information could come out in conversation, telephone clarification of reports, and other types of informal communicative processes. The data that are gathered through human observation are actually the essence of the HEAT system.

The HEAT methodology is therefore a viable example of a

collective performance measurement system which can qualify performance; identify training and procedural strengths and weaknesses; cumulate insights and lessons learned across exercises, scenario and organizations; and provide meaningful feedback to developers, operators, and trainers. Future research will be needed to ascertain the degree of success of this measurement tool.

J. Computer-Aided ARTEP Production System

Currently the Army, through the Army Research Institute (ARI) and Army Training Board (ATB), is developing a computer-aided ARTEP Production System (CAPS) to apply commercially available software to the ARTEP development and process, implement the system TRADOC-wide, and ultimately refine CAPS to take advantage of the wealth of information available from related automated systems and improvements in hardware and software (Meliza, 1986). The careful application of computerized technology to the ARTEP development process awaits demonstration; however, the concept holds promise for surmounting some of the problems present in current ARTEPs, such as lack of standardization in implementation and evaluation. It also may eventually capitalize on emerging methodologies, such as TEAM, HEAT, and clipboard technologies to advance measurement aspects of ARTEP.

K. Summary

Clearly, the development of computer capabilities has changed the content of unit performance measurement in the Army to a considerable extent. While still relying on subjective, decentralized processes of training and evaluation of combat effectiveness in the field through ARTEP systems, the Army's and other military services' trend toward more automated measurement of command behaviors through simulation poses a challenge to performance measurement. The complimentary nature of the two systems (automated and judgmental) opens a number of questions concerning the overlap of data acquisitions and analysis of similar information. For example, will a commander stationed in the field using ARTEP procedures utilize the same perceptions of field activities and arrive at the same decisions as a commander reacting to scenario-driven battlefield information presented off-site by computer? In addition, if these systems are to be diagnostic of deficiencies in training performance, will the feedback to participants and trainees be comparable, will one override the other, or will one system supplement the other in terms of detailed knowledge of results? These and other questions will need to be addressed as systems are further developed.

There is presently too much overload on the capabilities of commanders and evaluators in the ARTEP system. The expectation for them to develop evaluation plans, to train evaluators, and to prepare test documents clearly leaves too much to chance and to the expertise of the commander or evaluator and their training on the ARTEP system. This problem can subvert any reasonable attempt to establish reliability of measurement within the system. The findings of Thomas, Barber, and Kaplan (1984) that observers could not discriminate among ARTEP items, implying differing interpretations of ARTEP standards, reinforces this observation. On the other hand, in a related study (Thomas, 1984), objective indices produced by the CATTS system provided good correlations with battlefield performance.

There are obvious advantages to automated performance measurement systems. However, automated set up and review and other capabilities of computer-driven systems may also tax the information-processing and decision making capabilities of the observer/evaluator if output is not summarized and integrated in meaningful ways.

The HEAT system seems to provide the advantages of observational data coupled with exercise-generated data. In the final analysis, it will probably be a combination of computer-generated data and data generated from trained human observers that will provide the most accurate assessment of human performance in operational contexts. The challenge will be for researchers to establish the reliability of each form of assessment and to find that the two forms are not only correlated with each other but are also correlated with combat readiness, combat effectiveness or some other reliable criterion of training effectiveness.

Chapter V. Issues in Performance Measurement

There are a number of fundamental and important issues in performance measurement which have been alluded to in previous chapters describing the state-of-the-art in unit performance measurement methodologies but which have not been fully expanded. In this chapter we will focus on several of these issues as they relate to the development of sound performance measurement.

We stated in Chapter II that there are essentially three sources of measurement errors: (a) errors resulting from unreliability in the observation and recall of performance, (b) errors resulting from instability of the performance itself, and (c) errors associated with the deficiencies in the measurement instrument. We shall discuss each of these sources of error in turn and propose a number of guidelines or postulates that were developed during World War II by Robert J. Wherry, a psychometrician, to attack the issue of accurate and reliable measurement of task performance in improving military efficiency. Wherry devoted his efforts to developing a theory of rating and a series of postulates relating to the accuracy and reliability of performance ratings. Unfortunately , his work for the United States Army was never published in scientific journals which "was a tragedy since it offered a blueprint for research in this area". (Landy & Farr, 1983,p. 283). Landy and Farr (1983) fortunately published an appendix to their book on performance measurement, written by Wherry shortly before his death.

A. Errors Resulting from the Unreliability of Observation and Recall of Performance

Using human observation will usually cause the most influence on the measurement of the reliability. Lane (1986) states that humans, for a number of reasons, are probably incapable of recording events while excluding errors resulting from personal bias or preconceptions about the correct performance. When performing summary judgements, each observer has their own preconceived idea of the correctness of performance and this idea is integrated into their final judgement.

These limitations of the human observer are the reasons for the increased emphasis in the deriving of "objective" measurement systems. However, a total reliance on objective measures would eliminate performance measurement on some key areas such as decision making and other cognitive skills. As Vreuls

and Obermayer (1985) note, performance measures depend on overt actions; internal processes, such as decision making, however, may be manifest by simple actions or no actions at all. Automated performance measurement cannot make inferences about human processes, but human judgement can.

Many critics of current measurement systems equate "objective" with good performance measurement and "subjective" with bad performance measurement. Much of the rationale of these perceptions is based upon the presumption that reliability must out of necessity be increased by removing "measurement error" due to observers and replacing them with objective quantities that can be accurately recorded. However, Lane (1986) states that many studies show that "subjective" is not necessarily inferior and that "objective" is not necessarily superior. Furthermore, human observers and objective instrumentation are not necessarily measuring the same performance attributes. Muckler (1977) asserts that all objective measurement does involve some subjective judgements. The decisions about which physical measure will be recorded and how they will be translated into behavioral or performance measures are all subjective judgements. This subjectivity cannot be avoided and presents a problem for automated measurement systems because it concerns issues of "relevance" and "validity" of the measure set. Much of the literature suggests that human performance measures, while containing problems of bias and scale, are generally keyed to the detection of the appropriate aspects of performance (ie., they tend to be reasonably relevant and valid).

In addition, the deficiencies of human observations can be overcome by the pooling of judgements across observers and across time. Overcoming human observational weaknesses appears to be easier to surmcunt than choosing the wrong measures to be included in the objective measure set. In this case, including irrelevant measures will never result in a truly reliable performance measure (Lane, 1986).

Like Vreuls and Obermayer (1985), Lane (1986) observes that many aspects of performance are difficult to objectively measure, such as planning and "headwork". These behaviors are not well reflected in simple observations of inputs and outputs. Since these measures are important in performance measurement, their omission, however, would affect both the reliability and relevance of measures on the task. Danneskiold and Johnson (1954, cited in Lane, 1986) found that checklist-based measures were more reliable and produced higher correlations with other measures when these subjective judgements were combined with scores extracted from recorded observations. This notion has been suggested in the previous chapter.

In summary, the evidence seems to suggest that the more complex the skill for which measures are desired, the less likely it is that any single measure of performance will be reliable. Obtaining reliable measures for complex collective skills is likely to require the pooling of multiple estimates of performance, such as from both human observation and machine recording (cf., Martinko & Gardner, 1985).

The preceding considerations have been largely focussed on the measurement of individual performance. In discussing the problems associated with data collection and measurement in the field with teams of two or three members engaged in "real" work or training, Morgan et al. (1986) state that the challenges that must be met in order to obtain a clear picture of "what is going on " quickly become formidable. For example, they say, the experience described by Nieva et al., (1978) of Army units engaged in bridging a river, provides an illustration. In their research, practical obstacles led the researchers to resort to less refined descriptions and measurements than originally contemplated.

Likewise, in the Morgan et al. (1986) research, although the target-team members were in close proximity to one another and to the observers, it was not easy to keep track of what eight people (plus an instructor) engaged in a fast paced complex operation, were doing. The researchers therefore considered several ways to slow the pace or reduce the number of required observations. They considered videotaped transcriptions, the application of sampling strategies and the use of instructors and trainees as additional sources of information. In the end, they chose data sources to reflect all possible sources of information, including subject matter experts (instructors), trainees, and trained observers. Thus, complementary methods of data recording can come from multiple observers as well as combinations of human and non-human recorders.

In discussing the limits of what could be observed and recorded by available observers, Morgan et al. (1986) discovered distinct limitations with the use of video recording. The time, space, materials, people, and cost required to obtain satisfactory transcriptions seemed prohibitive for routine data collection purposes. However, the researchers suggested that video taping should be explored as a way to meet other objectives. For example., "Video recordings could be made of several teams as a resource for concentrated study of segments of special interest for research and training. Excerpts from the tapes could provide "live" examples for training instruction at the school and aboard ship". (Morgan et al., 1986, p. 69).

Many other training specialists (cf., Boldovici & Kramer,

1975; Goldstein, 1986) also endorse videotaping as having a great potential in the area of observer preparation. Films of exercises can be used to train observers on what to look for. After this training, observers make observations from films and then observations are correlated for inter-rater reliability. Videotaping thus makes the standardization of observers a less difficult task.

In regard to removing sources of unreliability due to errors in the observation and recall of performance, Wherry (in Landy & Farr, 1983) provides several relevant theorems:

Theorem. Raters will vary in the accuracy of ratings given in direct proportion to the relevancy of their previous contacts with the ratee.

Theorem. Raters will vary in the accuracy of ratings given in direct proportion to the number of previous relevant contacts with the ratee.

Theorem. Rating scale items that refer to easily observed behavior categories will result in more accurate ratings than those which refer to hard-to-observe behaviors.

Theorem. The rater will make more accurate ratings when forewarned about the types of activities to be rated, since this will facilitate the more proper focusing of attention on such pertinent behavior.

> Corollary a. Courses for the instruction of raters will be more efficient if they include instruction in what to look for.

Theorem. If the perceiver makes a conscious effort to be objective, after becoming aware of the biasing influence of a previous set, he or she may be able to reduce the influence of the bias.

> Corollary a. Training courses for the rater should include instruction about the effect of set on perception and provide practice in objectivity of observation.
>
> Corollary b. Deliberate direction of attention to the objective (measurable) results of behavior may serve to restrain the biasing effects of set.

Theorem. The keeping of a written record between rating peri-

ods of specifically observed critical incidents will improve the objectivity of recall.

Theorem. Any setting that facilitates the increase of bias, such as knowledge that the rating will have an immediate effect upon the recipient, will decrease the accuracy of raters, whereas any set that stresses the importance to the organization or to society as a whole will decrease perceived bias elements and thus increase accuracy.

Theorem. Knowledge that the rating given will have to be justified will serve unconsciously to affect the ratings given.

> Corollary a. Knowledge that the rating may have to be justified to the ratee may cause the rater to recall a higher proportion of favorable perceptions and thus lead to leniency.
>
> Corollary b. Knowledge that the rating may have to be justified to the rater's superior may cause the rater to recall a higher proportion of perceptions related to actions known to be of particular interest to the superior whether such actions are pertinent or not.
>
> Corollary c. To assure that neither of the distorting affects just mentioned shall take place alone, it is better to assure their mutual cancellation requiring that both types of review shall take place.

Theorem. Since forgetting is largely a function of intervening actions interposed between learning and recall, ratings secured soon after the observation period will be more accurate than those obtained after a considerable lapse of time.

Theorem. If observation is sufficiently frequent so as to constitute overlearning, the accuracy of recall will be improved.

Theorem. Observation with intention to remember will facilitate recall.

B. Errors Resulting From Instability of Performance Itself

The measurement of reliability can become difficult when the activity performed by the trainee consists of emerging behavioral skills (Lane, 1986). Skills that have not been learned adequately will usually be very unstable. Behaviors that are not stable cannot be measured reliably at a single point in time. Thorndike (1949) refers to these individual inconsistencies as intrinsic unreliability.

The measurement of reliability can also be influenced by random fluctuations in the task performance conditions. These fluctuations will result in unreliable performance measures even if the individual's performance is stable. Thorndike (1949) refers to these condition fluctuations as extrinsic unreliability.

The literature involving the analysis of reliability in aviation performance measures concerning both the stability of skilled behavior and the effects of changing task environments, has been very extensive. For example, Lintern, Westra, Nelson, Sheppard and Kennedy (1981) found that the reliabilities of the performance measures on simulators of aviation tasks were consistently low. The reliability of average approach scores in simulated carrier landings was about .38 . This low number was attained despite very precise data recording and reduction systems, which resulted in the measurement error being virtually zero; a large number of data points; and the precise control over the task conditions that is only possible in the simulation environment. Problems related to chronically low reliability in field measures have also been reported by Mixon (1982), Biers and Sauer (1982), and Johnson, Jones, and Kennedy (1984). Therefore, despite the use of objective measures and precise control over the changing task environment, the resulting reliability correlation can still be relatively small.

Lane (1986) reports that use of subjective measures for examination of the reliability of between-mission measures in aviation in almost every case resulted in reliabilities that were larger than or equivalent to the objective measures. For example, Crawford, Sollenberger, Ward, Brown, and Ghiselli (1947, cited in Lane, 1986) reported that the reliabilities of subjective criteria were "somewhat" higher.

In all of the studies comparing objective reliabilities to the subjective reliabilities, there emerge two consistent patterns: a) the day-to-day performance of an individual varies dramatically as a result of factors such as fatigue or changes in the manner of task performance, and b) the conditions in which the task is performed (weather, equipment, etc.) can actually account for more performance variability than the individual performance differences. Lane (1986) states that the superiority of subjective measures across successive performances can be explained; observers that provide summary judgements can to some extent take into account the effects of the varying task conditions and can make judgements relative to these conditions.

According to Lane (1986), the ultimate objective of all performance measurement systems is to quantify or assess in

some manner "good" or "proficient" performance in relation to a given task. Proficiency, it has also been discovered, is not necessarily the same as conformance to predescribed doctrine or standards. The cause of the variation is the many possible combinations of actions and reactions of the participants and participant's own judgements of different situations. The attempt to measure proficiency on a task or skill implies that the nature of the performance is fixed, in other words that it is an enduring characteristic of the individual being measured. However, Lane asserts that this idea of fixed performance is almost never true. For example, extended practice curves may effect the level of performance. Also, skilled performance tasks are often characterized by instability of initial performance and the presence of large individual differences in both the rate and shape of the acquisition curve. Only after the performance has stabilized can proficiency of the individual become a dependable measure. Measures prior to stabilization are really measures of progress and are not necessarily good predictors of ultimate proficiency for a given individual.

It is not difficult to see how such variabilities in individual performances are multiplied in situations where collective performance is to be assessed. The variability of group performance compounds measurement problems and suggests that performance should be monitored often over time to determine when performances reach stability.

Any team measure is made up of at least three separate components, representing a) the proficiency of individual team members on individual tasks, b) the proficiency of individual team members on their team tasks, and c) the learning curve resulting from continuing practice of the team as a unit, its evolving "cohesion". Each of these components contributes its own particular variance (true and error) to the collective team assessment. Most readily available indices of team performance are based on team output data: Such measures are unduly affected by the capability of the "weakest link" in the chain and may represent no more than the performance of that least able individual. They are also heavily dependent on events unrelated to "true" team performance (garbled transmissions, equipment malfunctions, etc.) and are sometimes directly determined by equipment driven pacing rather than by individual or team capability. Team output measures may thus not be usable because they contain large components of error and irrelevant variance (Turnage & Lane, 1987).

In addition, in the performance of complex tasks, many different components are required including perceptual skills, motor skills, planning, and the ability to make rapid and accurate decisions. Therefore, it is rather difficult to attain

one measure of proficient performance from all of these variables.

Ghiselli (1956, cited in Lane, 1986) reported three aspects of measure dimensionality. Static dimensionality denotes that at any one point in time, performance can be evaluated on a number of task dimensions. Dynamic dimensionality denotes that the dimensions important to success change across time. Individual dimensionality occurs when individuals are judged equally "effective" at doing the task, but differ in the components of the task emphasized to achieve results. These aspects of dimensionality further complicate collective performance measurement.

In regard to removing sources of unreliability due to errors resulting from instability of performance itself, Wherry (Landy & Farr, 1983) suggests the following theorems:

Theorem. Tasks in which the performance is maximally controlled by the ratee rather than by the work situation will be more favorable to accurate ratings. Any change of performance not due to the abilities of the ratee will only cause a source of error in the final rating.

Theorem. Rating scales or items that have as their behavioral referents those tasks that are maximally controlled by the ratee will lead to more accurate ratings than those that refer to tasks controlled by the work situation.

These theorems reiterate the necessity to separate individual and team task performances that are determined by ratee capabilities from those that are determined by external events.

C. Errors Associated With Deficiencies in the Measurement Instrument

The third source of measurement error in performance evaluation concerns variations in the measuring instrument itself and includes the very major problem of defining precisely what it is one wants to measure. The "criterion problem" is one that has consistently plagued researchers and does not lead to easy solutions. First, however, we will briefly discuss general scaling issues and formatting concerns.

1. Performance Rating Scales

Despite the fact that accurate performance rating scales are one of the most important parts of performance evaluation, present scalar systems are far from perfect. For the past 60 years, researchers have studied aspects of task rating in an

attempt to refine and perfect the accuracy of performance measurement.

One of the problems in performance measurement is the number of variables involved in the rating of tasks. Landy and Farr (1983) identify five different classes of variables that may be considered when examining performance appraisal: (a) the roles (rater and ratee), (b) the vehicle, (c) the rating context, (d) the rating process, and (e) the results of the rating. Appendix D presents an analysis of the vehicle, indicating the historical development of alternate rating scales over time.

Overall, the development of rating scales has sought to base evaluations on objectively defined, concrete behavioral observations. Efforts have been made to assure dimension independence (i.e., make sure that each behavioral incident relates to one and only one rating dimension), to reduce areas of rating error such as leniency, "halo", and central tendency, to involve users in the development of scales to improve content validity, and to improve inter-rater reliability. While scalar improvements have been made, we are still far away from the "perfect" instrument. Improvement efforts are likely to be constrained by issues concerning the purpose of measurement (description or evaluation) and the nature of the criterion.

2. Criterion-Referenced Versus Norm-Referenced Measures

Distinctions have been made between a) "norm-referenced" measures, from which ratings are developed by comparison to the measured performance of other individuals, and b) "criterion-referenced" measures, that are defined in terms of how a participant performs in relation to some established standard or "target value".

In criterion-referenced measures, such as ARTEP, the standards of "criteria" can be derived in at least three different ways (Lane, 1986). The criteria can be derived from some "natural point," such as a target (i.e., a target in a bombing exercise). However, this type of measure is notoriously unreliable, mostly due to the fact that they are outcome measures, which are influenced in operational settings by uncontrolled environmental changes.

The second method of deriving criteria involves the a prior definition of some "book" value. The attainment of this value is considered "good" performance.

The third method of establishing criteria is to determine empirically the behavior of "experts" in performing the tasks

or maneuver to create a "template" or profile, correspondence to which is defined as good performance.

Swezey (1978) summarized several different issues of criterion-referenced reliability. The principal problem is that traditional methods of computing reliability assume a true-score model. (i.e., all observations of performance are normally distributed around a "true" score). However, the scores that are achieved at the mastery level as in criterion-referenced measurement, are always at or near the ceiling value, and thus the variance and midpoints are correspondingly artificially restricted.

3. The "Criterion Problem"

All performance measures should contain the following three important characteristics: "1) Scales of measurement must be representative of, and capable of being directly mapped into, the "universe of events" that are ultimately important in successful outcomes for the task; 2) the scores assigned to individuals must be at least monotonic with respect to degrees of goodness/badness of the measured skill (s); and 3) differences among scores should be due primarily to differences in occurrences of "successful" events of processes rather than to other factors" (Lane, 1986, p. 61).

Thorndike (1949) coined the concept of the "ultimate criterion" which is a concept that embodies everything required for successful performance. Construct validity is the correlation between the measure and the ultimate criterion; however, such a measure is not very attainable.

Zedeck and Cascio (1984) described deficiencies in criterion theory as the major problem in performance measurement. Ash and Kroeker (1975) also described the limited progress made in criterion development over the previous two decades and suggested that further breakthroughs are not likely because of the complexity of the issue and the difficulty of experimentation in applied setting.

However, Waag and Knoop (1977, cited in Lane, 1986) and Breidenbach Ciavarelli, and Sievers (1985) provide the operations that are required in developing a "valid" measure set. A validation approach must contain as a minimum the following steps (Lane, 1986):

1. Identifying candidate measures. An excellent description of the procedures required for the development of the initial measure set are contained in Vreuls and Goldstein (1976) and by Vreuls (Obermayer, Woolridge

and Kelley (1985). Generally, the steps include a systematic analysis of the task, isolation of critical behaviors, and the determination of candidate parameters.

2. Reduction of the measure set. The measure set usually requires the elimination of irrelevant and unnecessary materials in order to attain a more realistic measure.

3. Selecting "valid" measures. The operations that link the candidate's measures to external variables (experience, subjective assessments, outcome measures, etc.) must have variance in common with the measure set to produce appropriate validity.

4. Determining the size of the measure set. The best way to reduce the number of variables is to use criteria external to the present data (e.g., expert judgement, perceived redundancy of content, or suspected unreliability).

In many cases, such as with ARTEP, choosing a criterion for measurement is determined by some "book" value that conveys the perception of validity in these measures (Lane, 1986). However, there exists a risk of "criterion deficiency" resulting from relying on the "book" value if in fact the "book" is not correct.

Subjectively-derived measures are another method for attaining the correct criterion for performance measurement. Observer ratings typically consist of a global rating that encompasses either overall proficiency or how well an individual performs a specific operation or maneuver. Lane (1986) states that the potential of completeness of subjective measures is relatively high, because of the ability of "experts" to combine judgementally a set of dimensions that are inherently different in meaning and on different scales. Hakel (1986, cited in Lane, 1986) suggests that the opinions of "experts" should be taken very seriously and considered a good source of performance information.

Furthermore, raters who are experienced in the tasks performed may differ in the relative weights assigned to the various aspects of performance, but are all probably keying on the "correct" aspects involved in the performance. These weights affect reliability and validity and diagnosticity, and therefore require a pooling of measures of the raters and occasions to "average out" rater biases. Many of the problems that have long been associated with subjective ratings can be eliminated by pooling the ratings, thus resulting in averages

that are relatively complete measures (Lane, 1986).

Combined measures also usually have a higher potential for "validity" than any of their components. This higher validity is a result of a better representation of the effects that influence the variation of the theoretical "true" performance, and have more variability in common with the "ultimate criterion" (Lane, 1986).

Despite the sound theoretical basis for maintaining separate measures, there still exist pressures for measurement systems that provide a single "overall" rating of proficiency. Thomas (1984) suggests that "unitary" measures are important for a number of reasons and are especially useful in the training setting. Unitary measures serve as general indicators for a) decisions about individuals, b) scaling the difficulty of training or practice to be given, c) evaluating the effectiveness of alternative training procedures, and d) general feedback to trainees.

The issue of whether and how to combine the separate components of multidimensional performance has been met with some controversy. Numerous writers have discussed this issue. For example, Dunnette (1963) asserts that a) the realm of performance is inherently multidimensional and should be viewed that way, b) there is no such thing as the one best criterion, and c) composite measures, despite the convenience and the appeal of their use, are unwarranted. These problems of skill instability and measure dimensionality have major implications in the selection and the reporting of results in performance measurement.

According to Lane (1986), although global measures are good for attaining the correct criterion, global measures are not useful in determining the reasons for a particular performance being deficient or proficient. If performance measures are to be used as guidance for improvement, then the variables contained in a measure set must be diagnostic. Measures of individual performance can be viewed as composed of two components, a) how well an individual understands what he needs to do, and b) his skill in execution of that understanding. Therefore, an important part of developing training criteria is the necessity for diagnostic procedures to determine specific improvement areas.

In order to ensure effective diagnostic use, the constructs estimated by the performance measures should be able to represent different aspects of skill, the obtained measures of those concepts should not correlate too highly, and each skill construct should be directly linked to some distinct score

(Lane, 1986).

One must also keep in mind that the validation process of a measure is only good in the context in which it was measured. For example, measures validated on the basis of group differences should be used on groups, but not as an assessment technique for individual deficiencies.

To overcome sources of unreliability associated with deficiencies in the measurement instrument , Wherry (in Landy & Farr, 1983) offers the following theorems:

Theorem. If the perceiver is furnished an easily accessible checklist of objective cues for the evaluation of performance, which can be referred to frequently, the perceiver should be better able to focus his or her attention properly.

Theorem. Physical features of a scale that facilitate recall of the actual perception of behavior will increase the accuracy of ratings.

Corollary a. Longer, objective, descriptive statements will be more effective than single value words or simple phrases in defining the steps on an adjectival type rating scale.

Corollary b. Overall ratings made after completion of a previous objective review (such as would be provided by the previous filling out of a checklist or forced-choice form) will be more accurate than those made without such a review.

Corollary c. The clearer (more self-explanatory) and more unambiguous the scale to be rated, the more likely that attention will be centered on the desired behavior.

Theorem. Performances that are readily classified by the observer into a special category will have relatively smaller overall bias components.

Corollary a. Jobs with simplified performance units requiring a single discrete aptitude will be rated with less overall bias than will complex jobs requiring a complex pattern of aptitudes.

Theorem. Rating items that are readily classified by the rater as referring to a given area of behavior will result in less overall bias than will items that suggest a complex pattern of behavior to the rater.

Theorem. The addition of extra qualified raters, with identi-

cal irrelevant contacts with a ratee, on a single item produces the same effect as the addition of extra items.

Theorem. The addition of enough extra qualified raters each with a completely different set of irrelevant contacts with the ratees will result in obtaining virtually true ratings.

Theorem. The addition of extra items of each type to the items of a heterogeneous scale will reduce error variance.

Theorem. To the extent that irrelevant rater contact with the ratees are somewhat different, the use of plural raters on a completely heterogeneous list of items will result in a reduction of bias.

Theorem. The addition of several extra items to each area of a completely heterogeneous scale to be used by several raters will further reduce error.

Theorem. The use of several raters on a scale composed of several items in each of several areas will further reduce error.

Theorem. The reliability of a rating item will be higher when determined by the test-retest, same rater method than when tested by the test-retest, different-rater method, and this superiority will increase as the difference in irrelevant (nonwork) contacts of the raters increases.

Theorem. The reliability of a single-item rating scale will be greater by the test-retest, same-rater method than when calculated by the item-alternate item, same-rater method.

Theorem. Halo will not disappear even when different raters are used until the irrelevant contacts with the ratees of the different raters is completely without overlap.

Theorem. A multi-item unidimensional scale is more reliable than a single item, but the relative proportion of true score to bias is not increased.

Theorem. Addition of several extra items to each area of a multidimensional test will increase reliability by decreasing error.

Thus, there are numerous steps which can be taken to increase the reliability of performance measurement when using human evaluators to rate behavior. All of these steps, exemplified by the foregoing theorems, are used to reduce sources of error and objectify ratings, thus becoming more descriptive

of "true" performance. An alternate method to objectify performance data, which is discussed in the next section, quantifies behavior by automated measurement systems.

D. Automated Performance Measurement

In the 1970's the introduction of the computer aided in improving performance measurement systems. The computer added better methods of both data recording and data reduction. Indeed, automated human-system performance measurement subsystems are now being specified as a requirement in modern training simulators (Vreuls & Obermayer, 1985).

One of the first methods to take advantage of the computer technology was criterion-referenced measurement in which system variables and participant actions taken are recorded at various points and then compared algorithmically to "tolerance bands" or other predifined standards of good performance. Good performance is defined in any criterion-referenced test as performing the job in a prescribed manner, according to specified procedures or doctrine.

An outgrowth of the technology that allowed the automated algorithm is the feasibility to examine the "process" by which an operator arrives at an end outcome or "product." The need for process measurement has been identified for many years but not until now has the technology enabled these measurements. Much of the field of individual skill acquisition and learning is better measured by the processes used to reach the acceptable outcomes than just a single outcome measurement.

Vreuls, Obermayer, Wooldridge and Kelly (1985) distinguish between the two concepts of assessment and quantification. They define assessment as requiring the use of many sources of information to determine the quality of performance, such as goodness or badness in relation to the specified training criteria. They also include in the definition of assessment the "evaluative" component of the measure set. The definition of measurement, on the other hand, includes operations which take deviations from the desired parameter values of performance.

According to Lane (1986), the complexity of arriving at final assessments has apparently discouraged such a final step in performance measurement. Semple, Cotton and Sullivan (1981), in a review of automated performance measures (APM's) on simulators, report the need to distinguish between true APM's and systems for automated data collection and recording. Semple et al. conclude that the present automated measurement capabilities in existing simulators "...are best described as performance monitoring or data collection systems" and that

"outputs from parameter monitoring capabilities are frequently not used" (p. 76).

Furthermore, Semple and Cross (1982, cited in Lane, 1986) suggest that such complex data recording systems are of low utility, because the volume of data produced is often overwhelming and can often be difficult to integrate and interpret.

Vreuls and Obermayer (1985) point out similar difficulties with automated measurement technology which are largely due to origins emerging from the use of automated methods for research. In such contexts, it may take long periods of time to clean up the large amount of recorded data, process wanted segments, and reach conclusions regarding those retained segments that show differences in experimental conditions.

Vreuls and Obermayer (1985) have indicated that there are many unanswered questions about the design of real-time automated measurement systems. Fundamental performance measurement problems in simulation relate to the hidden and embedded nature of performance, the lack of a general theory of human performance, determining validity of performance measures, and establishing criteria for performance. The latter two of these problems will be discussed later.

Vreuls and Obermeyer (1985) also point out some of the obstacles in implementing acceptable automated performance measurement systems. One of the most important is the difficulty of assessing complex human performance with anything approaching the expertise of human judgement. The development of measurement for performance diagnosis is another challenge because composite summary information measures are not likely to provide diagnostic information for analyzing performance problems. In addition, although complex systems are composed of teams of individuals who have designated functions, the contribution of each person to a team performance effort is difficult to define and measure. Adequate measurement may demand, for example, speech-understanding systems that are beyond the state of the art. Finally, Vreuls and Obermayer state that some tasks such as military field exercises, may not be amenable to automated measurement. For example, in communications or visual scan of displays, subtle cues operate so that observers frequently are required to capture the performance of interest as objectively as possible, a function APM cannot fulfill at present. Perhaps artificial intelligence (AI) systems can help in the future.

Despite these problems, Lane (1986) concludes that the advantages of simulator measurement have been well documented throughout the years. Some of these advantages are that envi-

ronmental and task conditions can be controlled, target behavior can be standardized, and scenarios can be repeated if desired. A properly developed simulator measurement system can be highly effective in the evaluation of trainee progress, particularly if it includes human observers in addition to APM.

E. Measuring Process Outcomes

One continuing criticism regarding the ARTEP system is its reliance on "product" rather than "process". That is why newer measurement systems such as the TEAM methodologies (Morgan et al., 1986) and the HEAT system (Defense Systems, Inc., 1984) focus on the processes involved in team performance.

Lane (1986) suggests that the ability to study the processes involved in producing an outcome is, in his judgement, the most important development in performance measurement over the last two decades. Outcome measures can be insensitive (and frequently inappropriate) indicators of the actual capability of participants. Measurement should therefore include (a) the manner in which outcomes are arrived at and (b) quantifiable performance of ability measures on the task components that account for the variance in those outcomes. Excluding these two components will lead to measures that will neither be diagnostic of performance nor useful in estimating the robustness of performance.

However, process variables have some practical disadvantages as performance variables. Sometimes measurement systems that rely heavily on operator judgements are not well understood and are not always well-suited to a process-type measure (Lane, 1986). Another problem that can complicate the rating of process measures is that no two people use exactly the same process to perform a specific task.

Thus, both process and outcome measurement is necessary. Trainers need to know both the "what" and the "why" of training performance. The concepts of ARTEP, TEAM, and HEAT methodologies and their associated advantages and disadvantages may provide a needed complement of methods and goals. It would be foolhardy to expect any one system to provide all the data necessary to simultaneously evaluate outcomes and diagnose deficiencies. At the same time, we need to know how these differing systems correspond to achieve improvement in training methods.

F. The Use of Proxy and Surrogate Measures

In many measurement situations, there exists a whole host of factors, both controllable and uncontrollable, that can have an influence on the final outcome. Some of these variables can be isolated and identified, while others can remain undetected despite very thorough analysis. When specific data on underlying factors is impractical to obtain, or the effects of such factors can not be isolated, then these internal variables sometimes serve as "proxy" measures. A proxy measure is a single quantity that reflects the combined effects of all the important determinants of performance of a task (Lane, 1986).

Lane (1986) also states that performance measurement sometimes, despite careful efforts, cannot easily be attained. Operational or field environments are especially susceptible to the problem of accurate performance measures. Operation or field environments may involve a phenomenon of interest that cannot be accurately measured in one assessment, because of its instability. Such is the case with combat performance. In this case, the nature of the task or the cost of each performance eliminates the possibility of repeated performances. In such a case, a surrogate measure could be used to replace the actual task. A surrogate measure is related to or is predictive of a construct of interest (i.e., "true" field performance), but is not a direct measure of that construct.

The concept of the "surrogate measure" has been developed for use in situations when operational measures cannot be measured with acceptable reliability. Surrogates must contain the four following characteristics: (a) correlate reasonably well to the performance construct, (b) be sensitive to the same factors that affect the unattainable performance, i.e., change in the same way in response to varying conditions as the performance variable would if it were accessible, (c) be more reliable than the field measures, and (d) involve minimal time to learn so that they can be used without extensive practice.

According to Lane, Kennedy, and Jones (1986), surrogates differ from other measure substitutes, such as those involving "synthetic tasks" or controlled job-sample approaches, in that the surrogate tasks take little time to learn, thus reducing the practice effects of repeated measures. Surrogates are also usually easier to score than other synthetic tasks. Thus, the concept of surrogate measurement may warrant future research interest.

G. Summary

The obstacles to reliable performance measurement are indeed formidable. First, individual performance is seldom stable, even when practiced over long periods of time on reliable, stable equipment and measured using objective recording procedures. When individual performance is complicated by interactions with others in unstable environments, measurement accuracy is similarly confounded. Second, although numerous strides have been made to train observers of performance to attend to relevant behaviors and to record them immediately without the interference of subjective evaluations, practical obstacles, or time delays, there are distinct limitations to observing behavior accurately. For this reason, it is desirable to increase the possible sources of information to include peer and self ratings as well as automated performance measurement, such as videotaping. The third source of measurement reliability, associated with the measurement instrument, involves a number of fundamental psychometric issues. What type of scaling format is best? Behaviorally anchored rating scales, based on critical incidents, are best overall although they involve complexities of dimension determination and are time consuming to construct. Furthermore, research (Turnage & Muchinsky, 1982) indicates that even well-trained observers fail to discriminate among dimensions. Should measures be criterion-referenced or norm-referenced? Criterion-referenced tests artificially restrict variance and do not lend themselves to classic psychometric evaluation. The "criterion" problem involves numerous decisions regarding how best to define and measure performances that come as close as possible to the "ultimate criterion". Should a unitary value be used or is a composite score preferable? Should observers be allowed to exercise judgement in evaluating performance or should they simply record neutral observations? What place does automated performance measurement have in emergent, dynamic, interactive, organismic team or collective performances? Should measurement focus on process or outcome? These questions, in the long run, are meaningless and unanswerable; they depend on the purpose of measurement and need not be artificially restricted by opting for one system over the other. If one follows the guidelines so well explicated by Wherry, then measurement accuracy can be increased. These guidelines stress control of extraneous sources of variance in performance of both the rater and the ratee. In spite of concerted efforts to control all sources of variance that contribute to measurement unreliability, however, most measurement systems as currently implemented fall short of the "ideal".

In these cases, "surrogate" measurement may be the alternative. The next chapter will summarize the current state of

performance measurement methodologies and suggest several avenues to pursue to improve the current state of affairs, including the use of surrogate measures.

Chapter VI. Summary, Research Implications, and Conclusions

A. Summary of Needs

This review has covered current team training research and unit training as it exists in military field operations. The focus has been on the methodologies used to measure collective performance with particular emphasis on the ARTEP system. An in-depth analysis of critical performance measurement issues emphasized the necessity to attend to basic reliability considerations to remove sources of error from the performance itself, the observation of performance, and the measurement instrument. Despite efforts to improve team training measurement systems, as exemplified by revised ARTEP, SIMNET, TEAM, and HEAT methodologies, the state-of-the-art of unit performance training and evaluation has remained at a fairly primitive level.

First, lack of understanding of some of the important dimensions of collective training and evaluation has handicapped the ability of trainers to recognize the critical components of successful performance in teams, squads, crews, and larger units. The need exists to develop acceptable definitions and models of collective performance and to differentiate individual, multi-individual, and collective skills and requirements. Methods are needed which can be used to identify individual and collective skill training requirements by using job/task analyses. The use of critical incidents methodologies to discriminate between isolated and interactive behaviors as they relate to overall mission success or failure would be one viable way to fulfill this need.

Second, in regard to developing collective training systems, any approach should permit the identification of the interacting communication, coordination, and decision making and other activities required in the task performance of each team member. ARTEP, TEAM, and HEAT systems provide examples of observational methods by which to identify and assess such behaviors, where SIMNET, COLTSIM, and SIMCAT offer examples of computerized simulation for multiple squad or platoon level elements. The identification of these behaviors is important because the individual must know his particular job as well as how he fits into the overall process of mission accomplishment. Thus, the ability to extract information on team process in a meaningful fashion is important to diagnostic team assessment and effective team training. Current assessment, however, is hampered because many military units frequently perform

qualitatively and quantitatively different missions and presumably perform them differently during peacetime than in wartime. The use of critical incidents methodologies to identify the fundamental characteristics of effective collective behaviors could provide standardized quantitative measures over quantifiable tasks.

Third, as we have pointed out in this review, the lack of adequate collective performance assessment methodology has resisted solution for the past 30 years and has resulted in an inability to provide needed information on the state of unit readiness and cohesion. The problems in observing team member interactions, coupled with the costs in money and time to conduct studies in this area and the unclear relationships between training variables and unit performance, have all contributed to the lack of success in seeking adequate measures of a unit's collective proficiency. The production of standardized, relatively invariant test conditions for evaluating dynamic, interactive team behaviors is one of the major problems in collective performance measurement. Other problems include the measurement problem of what to measure, where, when and how; these are still unanswered questions in the evaluation process. In addition, a criterion problem exists; without agreed-upon criteria or standards of collective performance, training proficiency cannot be evaluated. Thus, despite the many guidelines given us in previous chapters regarding procedures for improving the reliability of observer ratings through the application of rater training and improved measurement instrumentation, critical problems remain.

To further complicate matters, any attempt to create workable collective training and performance evaluation methods is challenged by inherently variable conditions of combat: mission, enemy, troops, terrain, time, and weather. Measurement considerations also must deal with the difficulty of observing the behavior of individual leaders/soldiers and teams. The number of activities occurring even in small teams and the rapidity with which they occur make thorough data collection impossible. The use of multiple observers helps to track team behaviors. The use of video equipment also helps, although equipment can often prove cumbersome for mobility. Critical incidents, as indicated before, provide a viable approach to sampling critical behaviors, but these incidents need to be extracted from a relatively complete record of activities. In addition, the actual presence of observers during training missions can cause reactivity and general distraction, thus affecting accurate data collection. Perhaps, as we shall address later, a combination of unobtrusive observations coupled with automated performance measures might overcome some of these difficulties.

The prevalence of these and other such complications to collective measurement suggests a need for innovative measurement approaches designed to address some of these chronic methodological problems, in particular those which might limit the reliability of measures and their sensitivity to conditions which might cause team performance to improve or degrade.

B. Suggestions for Measurement Improvement

The Army and other services recognize that research needs involve provision of innovative, low-cost solutions to long-standing methodological problems. One such solution, which has been described briefly in Chapter V, involves the use of surrogate measurement.

As Lane (1986) notes, it is often impossible to obtain satisfactory measures of performance because of the nature of the setting in which measurements must be taken (e.g., in operational or field environments). The preceding chapters have documented mechanisms that can lead to very low reliability in operational measurement. For example, there is extensive evidence that military operational performance measures lack reliability. Mixon (1982) and Lane (1986) summarized findings from pre-1985 literature and found chronically low reliability of measurements. These findings influenced the focus of a series of studies concerning reliability of criteria and of simulator performance on the Visual Technology Research Simulator (e.g., Lintern, Nelson, Sheppard, Westra, & Kennedy, 1981; Westra et al., 1986).

In these studies, single carrier landing approach performances on the simulator had test-retest reliabilities of .23 to .32; air-to-ground bomb miss distance reliabilities were somewhat lower, slightly above .20 . These low values occurred despite sophisticated and precise data acquisition systems and the ability to hold environmental and other variables constant (Lane, Kennedy, & Jones, 1986). Similar problems of unreliability were also encountered with field measures (Biers & Sauer, 1984; Johnson, Jones & Kennedy, 1984) In summary, Lane (1986) notes that a reliability of .30 is high for field measures, and .00 to .10 is typical for a single individual performance, leaving at least 90 percent of the field measure that cannot be related to anything else because it does not relate to itself. The case for team measurement reliability may be even lower.

Lane, Kennedy, & Jones (1986) have proposed the use of surrogate measure systems to overcome unreliability in op-

erational measures. The logic rests in the well-known correction for attenuation formula reported by Guilford (1954) and symbolized by:

$$R_t = \frac{r_{xy}}{\sqrt{r_{xx}\, r_{yy}}}$$

where r_{xy} is the predictive validity between the predictor variable (x) and the criterion variable (y), r_{xx} is the reliability of the predictor, r_{yy} is the reliability of the criterion, and R_t is the "true" relationship between perfectly reliable forms of x and y.

"Lack of reliability in either x or y constrains the possible relationship between the two. The impact of such an effect for operational performance measurement is very great. For example, if the field measure reliability is .40 and the predictor reliability is also .40, then an actual correlation of .40 is as high as one could obtain if all the reliable variance in the criterion were also shared with the predictor. More specifically, if the criterion is proficiency scores from maintenance tasks, and the predictor is hours of training, we might expect a true relationship of (say) $R_t = .75$; then the obtained correlation would be .30:

$$.75 = \frac{.30}{\sqrt{(.40)(.40)}}$$

Applying the attenuation formula often changes conclusions, and not always toward the negative. For instance, true predictive validities of operational criteria from paper-and-pencil aptitude tests are often underrepresented because of criterion unreliability" (Lane, Kennedy, & Jones, 1986).

A surrogate measure set is composed of tests or test batteries (usually simple "unidimensional" measures) specifically aimed at tapping the key components of a complex operational performance with a reliability high enough to compensate for any potential "irrelevancy" in the surrogate set. The basic metric appeal of surrogate measurement lies in the crossover between the reliability of a measure and its "validity" of criterion relevance. A field measure may have a reliability so low that it is mostly error, i.e., it has very little "true" performance variance. An external test (or battery), although it cannot be as "valid" as the measure itself, may tap more of the true variance of the field performance because its reliability is much greater. In general, when the correlation of a

test x is greater than the reliability of the field performance measure y ($r_{xx} \geq r_{yy}$), it is plausible to treat the test as a surrogate for the operational measure. In many cases, the costs of substituting a reliable and sensitive surrogate are much less than those required to attain sufficient improvement in field measures (Turnage & Lane, 1987).

According to Lane et al. (1986), "surrogates differ from intermediate or process measures taken during the performance of the task in that they are sensitive to the same factors as the criterion, but are entirely separate from the process of task performance itself. They differ from conventional performance measures in that the tests need not involve operations in common with the performance measures, only components or factors in common" (p. 1401). They also differ from other classes of measure substitutes, such as synthetic tasks (Alluisi & Fleischman, 1982) and controlled job-sample approaches (Biers & Sauer, 1982) in that little time is required to learn the task.

The development of surrogates entails demonstrating measurement properties through a series of operations similar to those used in establishing construct validity. A similar approach on an individual level is currently being used in the Joint-Service Job Performance Measurement (JPM) Project (Committee on the Performance of Military Personnel, 1986), although at this date it is too early to know the best combination of tests to serve as possible surrogates for measuring individual job performance (Kavanagh, Hedge, Borman, Vance, Kraiger, Dickinson, & Banks, 1987). Another similar approach is exemplified on the unit performance level by the Thomas, Barber, and Kaplan (1984) study of the impact of CATTS system characteristics on selected measures of battalion command group (BCG) performance. In that study, modified versions of two objective indices obtained from combat development studies were used as (surrogate) measures of battlefield performance: Relative Exchange Ratio (RER) and Surveying Maneuver Force Ratio Differential (SMFRD). These two measures were selected because they were found to be significantly correlated with the controllers rating of BCG performance: the change in combat ratio (CR) and command and control index of lethality levels (C2ILL).

C. Suggested Research

Based upon the preceding considerations, it should be possible to design a phased research program which would lead to significant improvements in the feasibility and effectiveness of a generic method for analyzing and assessing team performance.

A first step toward developing surrogate approaches to team measurement would be identification of the particular skills required by the team-specific components of task performance and the development of methods for testing these skills outside of the direct task context. Close observation of team performance and systematic isolation of the skills involved in communication, transmission of information, and team-paced procedures are essential to this process (Turnage & Lane, 1987).

Secondly, a surrogate methodology could be designed such that a number of existing measurement systems could be compared with each other. In a similar vein to the Thomas et al. (1984) study, subjective and objective procedures could be used simultaneously to comparatively assess the same team or collective performance. For example, ARTEP, TEAM, or HEAT systems could be used or modified to capture important "subjective" aspects of performance. At the same time, "objective" measurement involving computer simulation using SIMNET, COLTSIM, or SIMCAT systems, could provide relatively automated performance measurement. (These systems might be most appropriate to theater headquarters.) For example, HEAT methodologies (Defense Systems, Inc., 1984) are currently considering the design of the next generation of theater headquarters to take alternative structures into account. To effect this step, automatic data processing (ADP), linkages (communications) and size (manning) are viewed as determinants of effectiveness and the structure and functions have been established. This is a first step in integrating observational and automated data to determine "best predictors" or combinations of predictors. Comparisons of these records would provide a beginning to the analysis of team-specific behaviors and the measurement of those behaviors in a multitrait multimethod format. Systematic variations of training, environment, stressor conditions, and other factors which are likely to cause shifts in team performance can assist in determining the sensitivity of candidate surrogate sets to these manipulations.

One must not forget the analysis of actual on-the-ground team performance. For example, the National Training Center (NTC) at Fort Irwin and the new NTC for light forces at Fort Chaffee should provide a rich source of data on realistic though simulated combat performance. A surrogate measurement design might be developed within the context of the Army's Combined Arms Training Activity (CATA) using traditional ARTEP measurement techniques in conjunction with other measurement alternatives, including computer-driven battle simulations, MILES, radio/radar emitters, or robotics/AI.

Such a complex research design would, of course, need to be preceded by a distillation of candidate measures, through

interview, observation, literature review, or meta-analysis, where appropriate to identify the most reliable measures of performance. It should be noted that the Training Performance Data Center (TPDC), located in Orlando, Florida, could be the repository of many of these data sources including psychometric details on each measure and the effects of varying groups, environment, situations, and other combat-relevant variables. As in Project A (Kavanagh et al., 1987), the primary purpose would be to determine the psychometric properties of candidate measures under varying conditions, and if possible, over repeated measures. Reliable measures that showed significant correlations with criterion measures of battlefield effectiveness (and other criteria) could then be further evaluated for adherence to other prerequisites for adequacy of surrogate measurement. Significant findings would necessarily have to be crossed-validated. Thus the second phase of the research would involve validation of the methodology procedures and instruments considered viable in the first phase of the research, It would also involve revisions as necessary as well as inclusion of newly-discovered candidate measures. The final step of the research would be to field test reliable surrogate measures by applying techniques in the systematic design and development of experimental collective team training programs, by providing this training to selected teams, and by evaluating their performance.

D. Conclusion

The advancement of computerized automated simulation exercises such as JESS, BASE, ARTBASS, SIMNET, COLTSIM, and SIMCAT, has not been accompanied by advancements in measuring and evaluating command group performances in such a way that lessons can be learned from aggregating objective and subjective indices of critical performance. Although a new Combined Arms Lessons Learned (CALL) capability is established at the Combined Arms Training Activity (CATA) at Ft. Leavenworth SIMNET performance measurement technologies are advancing, further work is needed before quantifiable data bases are fully adequate to examine reliable correlates of critical combat performances with scientific rigor. Given the costs of fielding new training techniques, primarily simulation engagement exercises and equipment, it is imperative to make use of their inherent performance measurement capabilities. Indeed, new Manpower and Personnel Integration (MANPRINT) initiatives in the Department of the Army mandate development of unified, integrated MANPRINT data bases to define ranges of human performance in assessing new training devices.

Developments in emerging, objective automated performance measurement systems, combined with more precise quantification

in human observational systems, make the time right for comprehensive training program evaluation. For example, longitudinal evaluation programs that might examine the relationships among both objective and subjective measurement systems (e.g., APMs and ratings) as a function of varying tasks and conditions could suggest ways to focus on the "best" method to measure critical task performances. If such measures provided "good" data from a psychometric standpoint (i.e., were reliable and valid) over time, then trainers and evaluators would have a considerably easier time providing meaningful feedback to trainees.

The state of performance measurement in the military currently may require too much from trainers, observers, and evaluators. Trainers, particularly exercise commanders, are required to plan, coordinate, conduct, and critique mission exercises. Often, they simultaneously act as observers and evaluators. To observe and evaluate performance requires extensive military experience, and there is no assurance that the experience (and thus the evaluations) of one evaluator will necessarily correspond with those of another evaluator, despite requirements to adhere to doctrinal standards. When an evaluator has access to automated performance data, which is usually the case, the burdens of data integration and interpretation become excessive.

What is clearly needed is a way to simplify the task of individual evaluators to integrate and interpret data and, at the same time, provide a uniform data base from which multiple evaluators can draw similar conclusions. This ideal system, wherein reliable measures are input into a consistent and valid data-reduction process, can be approached at the present time if sufficiently proportionate funding is expended in performance measurement research as is expended on hardware and software engineering developments. Proper military training cannot be achieved by undue reliance on technological systems that provide questionable feedback; true combat readiness can only be achieved by the individual soldier knowing that his or her actions and the collective actions of the group have a high probability of success. The scientific study of performance, thus, can aid in the technical proficiency and motivational commitment of our military forces. In the long run, it is assumed that the more the soldier knows about critical individual and collective task performances, the more responsible he or she will become in performing those tasks.

Appendix A - Team Training Approaches in Other Services

A. Navy Team Training

There are five basic categories of Navy team training: Preteam Indoctrination Training; Subsystem Team Training; System Subteam Training; System Level Operational Training; and Multiunit System Operational Training (Wagner et al., 1977).

According to Morgan et al. (1986), the organization structure for naval combat includes integration of ship and support aircraft that defend against threats from the air, surface, and submarines. Each ship and aircraft has its own sensor and weapons operators and serves as components of larger "battle groups" to form complex interactive networks.

Training for battle group team members usually progresses from simulation-based instruction on individual operator tasks, through simulation training for subteams and single-platform teams, to simulation for multiple-platform teams. Finally, individuals are trained as a total battle group using operational equipment at sea, interspersed with additional training on shore-based simulators.

Much of the training resources go toward teaching members of battle groups how to work together to achieve common goals. For example, one training system is designed to combine anti-submarine and anti-surface warfare trains teams totalling over 10,000 personnel annually (Surface ASW, 1982, p. 1-1). Thus, most of the Navy training afloat is team training.

Again, the list of Naval team training devices is largely a review and update to Wagner's 1977 state-of-the-art review. They have been provided in this study largely for example purposes. The list is not intended to be comprehensive; it is more a guideline to the types of Navy team training available. Descriptions of the following devices: 2F87, 14A2, 14A6, 21A37/4, and TACDEW are directly from the Wagner et al., 1977 study. These devices are now out of date; therefore reviews of several new training devices are also included.

1. Device 2F87, Weapons Systems Trainer (WST)

The WST duplicates the interior arrangement and appearance of the P-3C aircraft. Five trainee stations simulate corresponding stations in the aircraft. The purpose of WST training is to teach team coordination. Students are organized into

teams according to position for which they are being trained. The trainees "fly" simulated anti-submarine warfare (ASW) missions. These missions are graduated in difficulty from very simple scenarios early in training to more complex exercises toward course completion.

2. Device 14A2, Surface Ship ASW (Anti-Submarine Warfare) Early Attack Weapons Systems Trainer

Device 14A2 is used to train surface ship teams in the proper utilization of operational ASW systems. Training emphasizes the procedural, tactical decision making, and team coordination activities in operating and employing ASW weapons systems. The device provides for indoctrination of personnel in ASW procedures and evaluation of tactical situations. The trainer is also used in developing and planning advanced Naval undersea defense tactics. The trainer occupies over 3000 square feet of floor space and duplicates the physical configuration of major operational compartments and equipments of surface ship Anti-Submarine Warfare (ASW) attack weapons. It simulates their functional operation and responses such as target detection, fire control solution, and weapon launching and tracking.

3. Device 14A6, ASW Coordinated Tactics Trainer

Device 14A6 is designed to train decision-making personnel in the tasks they must perform when engaged in coordinated ASW tactics. Simultaneous operation of 48 vehicles of various types and a multiplicity of sensors can be simulated. Communications facilities simulate the various radio channels employed operationally to coordinate all phases of an ASW mission from search through attack. Device 14A6 provides a synthetic environment within which ASW personnel can practice collecting and evaluating ASW information, making decisions, and implementing the decisions based on this information. The device is not intended to train equipment operators; therefore, simulated equipment are similar to fleet equipment. Virtually any exercise at sea which requires communication, coordination, maneuvering, and decision making may be practiced in the 14A6 trainer prior to going to sea.

4. Device 21A37/4, Submarine Fleet Ballistic Missile (FBM) Training Facility

Device 21A37/4 provides training in offensive and defensive tactics for nuclear attack center teams. Surface or subsurface maneuvers may be accomplished, and training may be given independently or in coordination with other units. In-

struction of senior command and staff officers in direction and coordination of submarine task groups with surface support units may also be given. A central digital computer provides problem generation, position and motion data generation. Up to 41 different vehicles can be included in training problems. A projection system in the attack centers permits both in-progress monitoring and post-fire analysis of training problems. Attack centers can be operated independently or operation can be coordinated to provide submarine versus submarine training. Fifteen different classifications of targets are currently available, 12 at any one time.

5. Tactical Advanced Combat Direction and Electronic Warfare System (TACDEW)

TACDEW is a highly sophisticated computer-based simulation facility having three primary missions: individual and team training, tactics evaluation, and testing of operational computer programs for Navy Tactical Data System (NTDS) ships. Training is conducted in 23 Combat Information Center (CIC) mock-ups typical of the ships on which trainees serve. Team and multiteam training are accomplished. The purpose of TACDEW training is not so much to establish team interactive skills as it is to maintain or enhance these skills in simulated mission contexts. Training typically consists of exercising a given team, or some combination of teams, in a series of scenarios of graded difficulty. The scenarios are designed to model tactical situations which might be encountered in an operational environment.

6. Surface AntiSubmarine Warfare (ASW) Training System and the Tactical System and the Tactical Team Training Device

These two training systems, which are currently under development by the Navy Training Systems Center (NAVTRASYSCEN), will purportedly cost $200 million to develop and $5 million per year to operate (Rees, 1985, cited in Morgan et al., 1986). The former system trains Combat Information Center (CIC), sonar, bridge and aircraft ASW operators for single-ship operations (Surface ASW, 1982, p. 2). The latter system extends this training to Anti-Surface Warfare (ASW) teams and emphasizes the coordination among and within ships and other platforms.

Additional team training requirements and costs come from the need to train at higher (Battle Force) command levels. At lower levels, separate team training is needed in areas such as air-to-air combat, air-to-ground combat, strike warfare (EW), casualty control, submarine diving maneuvers, naval gunfire

support operations, etc. (Morgan et al., 1986).

B. Marine Team Training

The Marine Corps is unlike the other services in that there is no formalized central training command although some support is provided by the Navy. Unit training and other training is monitored by the Commandant through the chain of command (Rosenblum, 1979). The Marine Corps places an emphasis on "Equip the Man," and "Man the Equipment." This focus is largely carried out through unit or team training. Unit or team training is in the hands of the commander at each level. The Marine Corps recognizes the need for computer-driven training systems in order to keep up with the state-of-the-art in unit training. Wagner (et al., 1977) describe the Tactical Exercises Simulator and Evaluator (TESE), and the Tactical Warfare Analysis and Evaluation System (TWAES). It should be noted that this information is once again quite dated and does not give sufficient credit to Marine initiatives in team training over the past 10 years.

1. Tactical Exercise Simulator and Evaluator (TESE)

The Tactical Exercise Simulator and Evaluator (TESE) is used to train Marine Corps officers in combat decision making. The project seeks to define procedures for war gaming. The goal is to get both computer-based individual and team measures during amphibious warfare exercises, and to increase the number of trainees who can be processed (Wagner et al., 1977).

2. Tactical Warfare Analysis and Evaluation System (TWAES)

The Tactical Warfare Analysis and Evaluation System is a computer-assisted system to control tactical field training exercises. The system offers potential improvements in maneuver control and simulation of indirect fire. The 1976 TWAES research effort studied the exact role that the post-exercise session would play in the total tactical exercise. The question being addressed was whether this feedback session would be a training vehicle to further extend the game's learning effectiveness or whether it would be merely a post-exercise debriefing during which administrative information about game procedures is passed on (Wagner et al., 1977).

C. Air Force Team Training

The Air Force uses a great number of simulators in their training. In fact, engagement simulation training plays a major role in Air Force training. However, since much Air Force

training is done on an individual basis, there are few training devices in the Air Force for team and unit training; the ones that do exist are largely for small man teams. Wagner et al., (1977) suggest that, with the lack of team training that exists, perhaps further training could be incorporated into more team training and development of team training devices, such as those listed below. Again, this list is out-of-date and is presented solely to illustrate the type of team training that the Air Force uses.

1. B52 (G&H Models) Weapons Systems Trainer (WST)

The WST integrates into a single team training device four individual devices: (a) The Mission Trainer used for training pilots and copilots; (b) Navigator Trainer used for training radar navigators and navigators; (c) Electronic Warfare Officer Trainer (EWO); and (d) Gunnery Trainer. The WST permits simultaneous team training of the entire six-man B52 crew (Wagner et al., 1977).

2. C5 Mission Flight Simulator

This team training device consists of three training stations which permit integrated team training for pilots/copilots, navigator and flight engineer (Wagner et al., 1977).

3. C130 Flight Simulator and C141 Flight Simulator

Both of these training devices are similar to the C5 Mission Flight Simulator except that training is not provided for the navigator station. However, the C130 Flight Simulator is scheduled to add the navigator station as part of the device in 1981 (Wagner et al., 1977).

4. Functional Integrated Systems Trainer (FIST)

The purpose of FIST is to provide a better means for training four members of the fire control team on the AC130E Gunship. The second objective of FIST is to refine and promote the use of the technology for developing low-cost, interlinked, functional, part task trainers (Wagner et al., 1977).

Appendix B - Other Army Performance Effectiveness Criteria

A. Inspection Scores Other Than ARTEP

Spencer et al. (1977) reported Army effectiveness measures other than ARTEP, the most common rating. They listed the second most commonly cited ratings in the Army evaluation inventory as IG inspection scores. The scores are from 0 to 2, with 0 being unsatisfactory, and 2 being outstanding. These scores are used on a variety of resources (i.e., personnel, equipment maintenance, cleanliness, etc.); however, their reliability is considered low due to the low variability of possible answers. Also, according to Spencer et al. (1977) inspectors are reluctant to give low scores (zeros) because of the detrimental effect to the commander.

The Annual General Inspection (AGI) is a more comprehensive evaluation of unit effectiveness which units go through annually. According to AR 20-3, these inspections "monitor the state of readiness, training, and mobilization throughout the Army" (Ulysses et al., 1983). The AGI, even though it touches on some non-combat related indicators of effectiveness, is primarily designed to measure unit combat effectiveness.

Technical Proficiency Inspection (TPI) scores are used for nuclear weapons and security. Their availability is obviously limited, but is available for qualified units at the battalion and installation levels. Their reliability is considered very good.

Computer-Operated Management Evaluation Technique (COMET) scores are the Army's quarterly inspections of all equipment (except TA50s--equipment issued to individuals). TA50 scores list equipment issued to individuals. The reliability and availability of these scores depends on the bookkeeping of each individual unit.

Finally, according to Spencer et al.(1977), Skill Qualifications Tests (SQTs) are considered a very good source if the test being used is practical and specific and not a paper and pencil examination. Also, if field days are held, then they produce a variety of information on unit, company, and battalion level activities. However, their availability seems to be the option of the unit and thus reliability would most likely have a wide variance.

B. Mission Accomplishment Results

Mission accomplishment results include number of hours flown in air units, accident rates, and mission objectives. The first two measures usually keep by strict records and therefore are available and reliable (Spencer et al., 1977)

If a unit has a "management by objectives" plan, then the number of objectives in that plan are the mission objectives. These are most likely to vary greatly as no two units are identical. Spencer et al. (1977) found that their availability was questionable in most units.

C. Efficiency Measures

1. Unit Status Reports

The Unit Status Report is the Army's major method for measuring unit effectiveness. A Unit Status Report is required for most units every month and is concerned with monitoring indicators of combat readiness. The three components of the Unit Status Report (equipment readiness, training, and personnel readiness) are supposed to provide an indication of a unit's overall combat readiness. However, the validity of this report has never been established in a scientific manner. Moreover, the reliability of the information provided by unit commanders on the Unit Status Report is, at times, inaccurate (Spencer et al., 1977).

2. Operational Readiness Status Reports ("OR rates")

This report gives the readiness status of each piece of equipment in the unit (operational, non-operational). It is available in some but not all units. It is considered of poor reliability (Spencer et al., 1977).

Ryan and Yates (1977) in a study of the face validity of the Operational Readiness Training Tests (ORTT's) found generally positive results, but recommendations for improvement were suggested to make the training more realistic with respect to how the enemy might behave in combat. However, without some form of control for the Operational Readiness Training Test, the positive result could be due to general cooperativeness on the part of respondents or other similar methodological artifacts. In addition, ORTT's are reportedly being phased out.

Thus, these combat effectiveness indicators are seldom accurate measures of the commander's performance. In addition, they lack a guide for unit commanders which identifies the best indicators and how to use them properly and they show no reli-

ability. These effectiveness measures are largely replaced by improved ARTEP evaluation and the use of automated performance measurement using simulators.

3. Other Efficiency Measures

Other efficiency measures include deadline reports (the estimated time of repair of a piece of equipment), equipment casualty reports, maintenance requests (Form 2404), parts requisitions, equipment lost reports, service requests, and cost budgets. The availability and reliability of all these measures are major problems in this area according to Spencer et al. (1977), largely due to problems in the quality of record keeping.

D. Personnel Development Measures

Finally, the Army evaluates individual performance in a number of different ways (Spencer et al., 1977).

Promotions are measured by the percentage of persons who are presently eligible for promotion, or who have been promoted. This score focuses on the more important rank levels. It is often used to compare with minority group performance. Spencer believes that it may be used indirectly as a measure of superior and unit training effectiveness. These data are largely available and reliable (Spencer et al., 1977).

Education indices come from the number or percentage of persons who are taking or have completed optional courses. Data are available for some units and reliability is good where records are kept (Spencer et al., 1977).

Physical training (PT) tests and awards, including unit citations and individual citations, are other forms of data that the Army uses to evaluate personnel development. These data vary in reliability as a function of how subjective and competitive the evaluations are.

Appendix C - Evaluation Systems Used in Other Services

A. Navy Evaluation Systems

The Navy makes use of many simulators and devices for unit training. The measurement strategies and problems associated with simulators and devices are largely the same for the Navy as for the Army SIMNET and other simulation systems. They too, require trained instructors to provide meaningful feedback and evaluation to trainees.

Many types of paperwork are used to evaluate the effectiveness of training and training equipment. The two major systems used by the Navy are the Operational Readiness Evaluation (ORE) and the Operational Readiness Inspection (ORI).

The Operational Readiness Evaluation is one of the Navy's most often used evaluation assessment instruments. This evaluation usually consists of evaluating: personnel, equipment, personal equipment, training, etc. It is similar the Army's IG Inspection, COMET, and TA50 scores. According to Spencer et. al. (1977), availability of data is good, with reliability depending on unit's record keeping.

The Operational Readiness Inspection is basically the same as the ORE for the Navy. It serves the same basic functions as the Operational Readiness Evaluations and includes the status of equipment and training, similar to the Army's and to the Marine Corps's evaluation systems.

B. Marine Corps Evaluation Systems

The Marine Corps has several major evaluation systems, with the Marine Corps Combat Readiness Evaluations System (MCCRES) being the most important. It serves as the base for the other systems described below and is comparable to the Army's ARTEPs. (Rosenblum, 1979).

According to Rosenblum (1979), "Marine Corps Combat Readiness Evaluations System (MCCRES) is a system which provides unit proficiency standards. These standards are the basis for training at battalion and division levels. Units are evaluated against these standards for readiness reporting purposes at the Fleet Marine Force (FMF) level.

The Tactical Warfare Simulation, Evaluation and Analys's

System (TWSEAS) is a portable system which complements MCCRES through the training of staffs. TWSEAS realistically games opposing force meeting engagements and includes contingency plans to evaluate the performance of the staff while executing them. Another evaluation system is the Marine Corps Key Evaluation System (MCKEES), which is a computerized data retrieval system. The data contains lessons learned and problem areas from previous Marine Corps operations, and includes MCCRES evaluations. It is designed to be integrated into TWSEAS and it also serves as a readily accessible source of information for commanders to study pitfalls for use in the planning of future operations.

The Marine Corps Air Ground Combat Center (MCAGCC) located at 29 Palms is an organization which trains and evaluates unit performance in air/ground and fire support coordination procedures in a realistic, live fire environment using appropriate MCKEES data and the TWSEAS gaming process. Thus, Marine unit training systems and evaluations of those systems are similar to those used by the Army and Navy.

C. Air Force Evaluation Systems

The Air Force uses the Operational Readiness Inspection measure as well as On-the-Job Training programs. While the other services would not deny the existence or occurrence of on-the-job training, they do not have full-time On-the-Job Training (OJT) Managers (Rosenblum, 1979).

The Air Force Operational Readiness Inspection (ORI) is the same as that of the Navy and is comparable to the Army's ARTEPs and the Marine Corps' MCCRES used to assess the readiness capabilities of the unit/units in question.

On-the-Job Training (OJT) occurs in all of the Armed Forces to some degree, but the Air Force seems to take it more seriously. They employ full-time On-the-Job Training Managers, who oversee all on-the-job training of recruits, commanders, and others. Rosenblum (1979) recommends that each service follow the Air Force's example and employ OJT Managers. He states that this would lead to better OJT programs and would help the unit commanders achieve higher standards of training.

Appendix D - Analysis of Alternate Rating Scales

A. Graphic Rating Scales

Graphic rating scales, introduced in 1922 by Paterson, consist of trait levels and brief explanations of those levels. The scale consists of an unbroken line with varying types and numbers of adjectives below. Some variations on the scale include segmenting the line and numbering the scale levels (Madden & Bourden, 1964). Graphic scales have most often appeared in research as a comparison to more recently developed scaling systems and represent a primitive, subjective form of rating.

B. Forced-Choice Scales

The forced-choice scale developed out of dissatisfaction with the conventional rating scales such as the graphic scale. The Army has used the forced-choice scale in the rating of officers (Kane, 1985). The basic format consists of statements carefully grouped together into pairs, triplets or tetrads. The rater is asked to choose either the most descriptive or least descriptive statement in relation to the ratee.

The advantage of the forced-choice format is that the statements grouped together appear to have equal value, but the statements within the group actually differ in how well they differentiate between successful and unsuccessful performance on a particular characteristic. This format reduces the subjective element in evaluations, and is resistant to fallibility (Zavala, 1965). However, the forced-choice scale has one major disadvantage; the technique tends to be unacceptable to raters because they have little control over the outcome of the appraisal (Baier, 1951).

When compared with the graphic scales, the forced-choice scales exhibit less leniency, defined as the bias of raters to be lenient in their ratings (Howe & Silverstein, 1960; Sharon & Bartlett, 1969; Taylor, Schneider & Clay, 1954). The forced-choice scale also combats the problem of range restriction where raters fail to use the full range of scores (Cotton & Stolz, 1960).

The reliability and validity of the forced-choice scale has been questioned. Isard (1956) found that a forced-choice scale of socially ambiguous statements was reliable and valid. However, Kay (1959) found low reliability and validity when using the forced-choice scale for critical incidents.

C. Behaviorally Anchored Rating Scales

The Behavioral Expectation Scale was introduced in 1963 by Smith and Kendall. This scale has also been labeled the Behaviorally Anchored Rating Scale (BARS). BARS are developed by collecting critical incidents of behavior and clustering them into different performance dimensions. The incidents are then independently reassigned to the dimensions by subject matter experts to ensure that the dimensions are both accurate and valid. Then the incidents are rated for each dimension according to how effective or ineffective the incident is judged to be by experts. The final instrument is formed by six or seven scaled incidents for each dimension (Schwab, Heneman, & DeCotiis, 1975). This is the type of scale used by Morgan et al., (1986) in their TEAM performance measurement program.

Schwab et al., (1975) point out several advantages of the BARS system. The participation of users in the scale development results in increased content-validity. This participation also results in better understanding and use of the instrument. Hom, DeNisi, Kinick, and Bannister (1982) report the participation of the users will also increase the effectiveness of the feedback given to the participant, because the scale dimensions are anchored in defined concrete behavioral terms.

Schwab et al., (1975) reported that the BARS system was superior to other formats in the areas of leniency, dimension independence and reliability, However some studies suggest little or no difference between BARS and the graphic rating scales in these areas. The conclusion of the research on the BARS system is that it has not performed up to expectations in the areas of leniency, dimension independence and inter-rater reliability. Kingstrom and Bass (1981) report that the empirical evidence suggests that BARS may be generally less susceptible to observed halo error (i.e., inability to discriminate among dimensions when making ratings) than other formats, although the differences are small and are seldom statistically significant. Kingstrom and Bass (1981) conclude that there are small (if any) differences between BARS instruments and other scale formats on various psychometric characteristics.

Several studies on the BARS system have demonstrated convergent validity (Borman, 1979; Dickinson & Zellinger, 1980) indicating overlapping results using different scale formats. A few studies have reported greater discriminant validity of dimensions using BARS (Arvey & Hoyle, 1974; Campbell, Dunnette, Arvey, & Hellervick, 1973) compared with other scale formats.

D. Behavioral Observation Scales

The Behavioral Observation Scale (BOS) was developed by Latham and Wexley (1977). First, scale items are developed by using the critical incident technique (Flanagan, 1954). Then similar items are grouped together to form one behavioral dimension, using job incumbents and/or analysts to form the dimensions. The incidents are then placed in random order and given to a second individual or group in order to reclassify the incidents according to the original categorization system. This process is very similar to the retranslation in BARS. A Likert scale is added to each behavioral item in order to indicate the frequency that the job incumbent has been observed performing the behavior.

The advantages of the BOS system are cited in Latham, Fay, and Saari (1979). The scales are based on user input which increases the content validity of the instrument. BOS is also based on behaviors that are performed which increases the value of the feedback. The behavioral based feedback is important because the participant can receive feedback in specific areas of needed improvement. Another advantage of BOS is that the raters are only required to make observations concerning the frequency of a given behavior. However, this claim is questioned by Murphy, Martin and Garcia (1982), who suggest that memory is not a simple compilation of events, but is affected by the rater's general impression on recall of specific events.

Borman (1979) assessed the validity and accuracy of five different rating formats using videotaped performers on two different jobs. He found that all of the formats were equal in convergent validity, but found a summed BOS scale exhibited slightly higher discriminant validity. But, Borman also found that the summed scale was the lowest on the measure of accuracy. In conclusion, studies have shown that the BOS format and summed scales generally perform at least as well as the other formats when assessed for degree of halo and leniency error.

E. Performance Distribution Assessment

The BOS method does, however, possess some inadequacies. One problem with BOS is its failure to explicitly focus on the quality or value of a given behavior. Another problem, is that BOS characterizes each person's performance in terms of a single point on a continuum of goodness. Kane and Lawler (1979) conclude that the rater's task becomes one of selecting a single goodness level to represent an entire performance distribution.

Kane (1985) developed an appraisal system designed to overcome these and other shortcomings of previous appraisal methods. Kane defines performance on a job function as "the record of outcomes achieved in carrying out the job function during a specified period" (p. 2). He further distinguishes job functions as either "iterated", meaning that they are carried out on two or more occasions during an appraisal period, or "non-iterated," meaning that the are carried out only once during an appraisal period. Kane reports that most job functions are iterated, and suggests that his appraisal method, Performance Distribution Assessment (PDA), is the most appropriate and effective method to measure performance on iterated job functions.

Performance for an iterated job function can be represented in a quantifiable manner on a performance distribution which represents the rate at which the performer achieved each specified outcome level for a given job function. Kane also asserts that no part of the performance variation is derived from random measurement error. He suggests that the variation of outcomes achieved is due to the performer's varying level of motivation and to the restrictive influence of extraneous constraints beyond the performer's control.

Kane also suggests that the human role in the measurement procedure should be reduced to an absolute minimum in order to combat the role of human error associated with the rating process. This can be accomplished by minimizing the number of steps in the measurement process assigned to the human role. In order to further reduce the human error, Kane suggests a highly specific set of rules to guide the rater in the measurement process. The rules should proceduralize the human role in the measurement process. Kane's PDA method attempts to fulfill these needs.

The PDA method becomes operational by specifying the job functions on which the participant is to be appraised and the criteria for each job function. Three levels of performance measurement outcomes are proposed for each job function criterion: the most effective outcome and an intermediate outcome. Two other outcomes are also indirectly defined as falling between these three. For each outcome level the rater is asked to specify: on what percentage of occasions did this person actually perform at this level or higher.

Kane (1986) concludes that the PDA is superior to other systems, because it excludes any portion of the range of performance that is impossible to achieve, due to circumstances beyond the participants' control.

The one disadvantage of the PDA method is that it is complicated and time consuming to score; however, the use of computers may be able to eliminate this problem. The methodology is currently being assessed in an operational context against a BOS instrument, so it is still premature to conclude whether Kane's method will indeed be superior to other systems.

References

Alluisi, E. A., (1977). Lessons from a study of defense training technology. Educational Technology Systems, 5 (1), 57-76.

Alluisi, E. A., & Fleischman, E. A. (Eds.) (1982). Human performance and productivity: Vol. 3: Stress and performance effectiveness. Hillsdale, N.J.: Erlbaum.

ARTEP FC7-13 Light Infantry Battalion ARTEP Mission Training Plan (AMTP). Washington, DC: U.S. Army Infantry School.

Arvey, R. D., & Hoyle, J. S. (1974). A Guttman approach to the development of behaviorally based rating scales for systems analysts and programmer/analysts. Journal of Applied Psychology, 59, 61-68.

Ash, P. & Kroeker, L. P. (1975). Personnel selection, classifi cation, and placement. Annual Review of Psychology, 20, 481-507.

Baier, D. E. (1951). Reply to A. Travers' "A critical review of the validity and rational of the forced-choice technique." Psychological Bulletin, 48, 421-434.

Barber, H. F., & Kaplan, I. T. (1979). Battalion command group performance in simulated combat (Technical Paper 353). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A070 089)

Battle Simulation Software Survey. (1986). Orlando, FL: U.S. Army Project Manager for Training Devices, Naval Training Center.

Bauer, R. W. (1981). Review of methodologies for analysis of collective tasks (Research Report 1329). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A130 971).

Baum, D., Modrick, J., & Hollingsworth, S. (1981). The status of Air Force team training for command and control systems (Report No. 81sRC14). Minneapolis, MN: Honeywell Systems and Research Center.

Biers, D. W., & Sauer, D. W. (1982). Job sample tests as predictors of M-1 gunnery performance (Appendices A-E) (Report No. SRL-6639-1). Dayton, OH: Systems Research Labs, Inc.

Boldovici, J. A. & Kraemer, R. E. (1975). Specifying and measuring unit performance objectives (Final Report DAHC19-73-C-0004). Alexandria, VA: Human Resources Research Organization (HumRRO).

Borman, W. C. (1979). Format and training effects on rating accuracy and rater errors. Journal of Applied Psychology, 64, 410-421.

Boldenbach, S. T., Ciavarelli, A. P., & Sievers, R. (1985). Methods and metrics for aircrew assessment during close-in air-to-air combat (R-87006). San Diego, CA: Cubic Corporation.

Briggs, G. E., & Johnston, W. A. (1965). Team training research (Report No. 1327-2). Orlando, FL: Naval Training Device Center.

Campbell, J. P., Dunnette, M. D., Arvey, R. D. & Hellervick, L. V. (1973). The development and evaluation of behaviorally based rating scales. Journal of Applied Psychology, 57, 15-22.

Chesler, D. S. (1971). Computer-assisted performance evaluation for Navy Anti-Air Warfare training: Concepts, methods, and constraints (Research Report NPTRL SRR71-25). San Diego, CA: Naval Personnel and Training Research Laboratory.

Committee on the Performance of Military Personnel (1986). Assessing the performance of enlisted personnel: Evaluation of a joint-service research project. Washington, DC: National Academy Press.

Connelly, E. M., Comeau, R. F. & Steinhauser, F. (1980). Team performance measures for computerized systems. (Final Technical Report, Contract No. MDA-903-79-C-0274, Conducted for Army Research Institute for the Behavioral and Social Sciences). Vienna, VA: Performance Measurement Associates.

Cooper, M., Shiflett, S., Korotkin, A. L. & Fleishman, E. A. (1984). Command and control teams: Techniques for assessing team performance (Final Report F33615-81-C-0017, AFHRL-TP-84-3). Bethesda, MD: Advanced Research Resources Organization.

Cotton, J. & Stolz, R. E. (1960). The general applicability of a scale for rating research productivity. Journal of Applied Psychology, 44, 276-277.

Daniels, R. W., Alden, D. G., Kanarick, A. F., Gray, T. H. & Feuge, R. L. (1972). Automated operator instruction in team tactics (Task No. 8505-1). Orlando, FL: Naval Training Device Center.

Dees, J. W. (1969). Squad performance as a function of the distribution of a squad radio (HumRRO TR 69-24). Alexandria, VA: Human Resources Research Organization. (AD A701 152)

Defense Systems, Inc. (1984). HEAT Executive Summary. Prepared for the Defense Communications Agency/PSI-CAMA.

Denson, R. W. (1981). Team training: Literature review and annotated bibliography (AFHRL-TR-80-40). Dayton OH: Logistic and Technical Training Division, Air Force Human Resources Laboratory, Wright-Patterson Air Force Base.

Dickinson, T. L. & Zellinger, P. M. (1980). A comparison of the behaviorally anchored rating and mixed standard scale formats. Journal of Applied Psychology, 65, 147-154.

Dunnette, M. D. (1963). A note on the criterion. Journal of Applied Psychology, 47, 225-251.

Dyer, J. L. (1986). Annotated bibliography and state-of-the-art review of the field of team training as it relates to military teams (ARI RN 86-18). Alexandria, VA: U.S. Army Research Institute for the Behavioral & Social Sciences. (AD A169 195)

Ericksen, S. C. 1952). A review of the literature on methods of measuring pilot proficiency (Research Bulletin 52-25). Lackland AFB, TX: Human Resources Research Center.

Finley, D. L., Rheinlander, T. W., Thompson, E. A. & Sullivan, D. J. (1972). Training effectiveness evaluation of Naval training devices, Part I: A study of the effectiveness of a carrier air traffic control center training device (Technical Report, NAVTRAEQUIPCEN 70-C-0258-1). Westlake Village, CA: Bunker-Ramo, Electronic Systems Division. (AD 146 414)

Flanagan, J. C. (1954). The critical incident technique. Psychological Bulletin, 51(4), pp. 327-358.

Glanzer, M. & Glaser, R. (1955). A review of team training problems. (Prepared for Office of Naval Research). Pittsburg, PA: American Institutes for Research. (AD A078 434)

Glickman, A. S. & Vallance, T. R. (1985). Curriculum assessment with critical incidents. Journal of Applied Psychology, 42, 329-335.

Goldin, S. E. & Thorndyke, P. W. (Eds). (1980). Improving team performance: Proceedings of the Rand team performance workshop (R-2606-ONR). Santa Monica, CA: Rand Corporation.

Goldstein, I. L. (1986). Training in organizations: Needs assessment, development, and evaluation: (2nd Ed.). Monterey, CA: Brooks/Cole Publishing Co.

Guilford, J. P. (1954). Psychometric methods (2nd Ed.). New York: McGraw Hill.

Hackman, R. J. (1982). A set of methods for research on work teams (Interim Report N00014-80-C-0555). New Haven, CT: Yale School of Organization and Management.

Hagan, C. (1981). "Improved ARTEP in Reforger 1981 and the Future," In Sulzen, R. H. Unit Performance Measurement: Proceedings of the ARI Sponsored Seminar. Alexandria, VA: U.S. Army Research Institute for the Behavioral & Social Sciences.

Hall, E. R. & Rizzo, W. A. (1975). An assessment of U.S. Navy tactical team training: Focus on the trained man (TAEG Report No. 18). Orlando, FL: Training Analysis and Evaluation Group.

Hammell, T. S. & Mara, T. D. (1970). Application of decision making and team training research to optimal training: A translative technique (NAVTRADEVCEN 68-C-0242-1). Groton, CT: General Dynamics.

Harris, D. A. (1986). Joint-service job performance measurement/ enlistment standards project. Personnel communication.

Havron, M. D., Gorham, W. A., Nordie, P. G. & Bradford, R. G. (1955). Tactical training of the infantry rifle squad (HumRRO Technical Report 19). Washington, DC: Psychological Research Associates, George Washington University.

Havron, M. D. Albert, D. D., McCullough, T. J., Johnson, E., III, & Wunschura, R. G. (1978). Improved ARTEP methods for unit evaluation. Volume I. Executive summary; Study design & field research (ARI Technical Report 78-A26). Alexandria, VA: U.S. Army Research Institute for the Behavioral & Social Sciences. (AD A064 271)

Havron, M. D. Albert, D. D., McCullough, T. J. & Wunschura, R. G. (1978). Improved ARTEP methods for unit evaluation: Volume II. Analysis (ARI Technical Report 78-A27). Alexandria, VA: U.S. Army Research Institute for the Behavioral & Social Sciences. (AD A066 783)

Havron, M. D., Albert, D. D., McCullough, T. J. & Wunschura, R. G. (1978). Improved ARTEP methods for unit evaluation: Volume III: Field guidance (Technical Report 78-A28). DAHC19-77-C0001). Alexandria, VA: U.S. Army Research Institute for the Behavioral & Social Sciences. (AD A064 272)

Havron, M. D., McCullough, T. J., McFarling, L. H. & Wunschura, R. G. (1979). Improved ARTEP methods for unit evaluation: Volume IV: Guidelines for planning and conduct of company level field exercises (Research Product 79-1). Alexandria, VA: U.S. Army Research Institute for the Behavioral & Social Sciences. (AD A075 470)

Havron, M. D., McFarling, L. H. & Wunschura, R. G. (1979). Improved ARTEP methods for unit evaluation: Volume V: Analysis of alternative training settings. (Technical Report 79-A23). Alexandria, VA: U.S. Army Research Institute for the Behavioral & Social Sciences. (AD A075 465)

Havron, M. D., McFarling, L. H., Hill, H. & Wunschura, R. G. (1979). Improved ARTEP methods for unit evaluation: Volume VI. Conventional ARTEP missions and engagement simulations: Examination of options (Technical Report 79-A24). Alexandria, VA: U.S. Army Research Institute for the Behavioral & Social Sciences. (AD A075 663)

Havron, M. D. & Wunschura, R. G. (1979). Improved ARTEP methods for unit evaluation: Volume VII: Executive summary (Technical Report 79-A25). Alexandria, VA: U.S. Army Research Institute for the Behavioral & Social Sciences. (AD A076 957)

Hayes, R. E., Davis, P. C., Hayes, J. J., Abolfathi, F. Harvey, B. & Kenyon, G. (1977). Measurement of combat effectiveness in Marine Corps infantry battalions. Arlington, VA: Defense Advanced Research Projects Agency.

Herzog, L. A. (1986). Simulation networking--SIMNET. Proceedings Armor/Calvary Force in the 1980-90s Conference. Fort Knox, KY: U.S. Army Armor Center.

Hollenbeck, N. A. (1986). COLTSIM, company/team level tactical simulator. Proceedings Armor/Calvary Force in the 1980-90s Conference. Fort Knox, KY: U.S. Army Armor Center.

Hom, P. W., DeNisi, A. S., Kinick, A. J. & Bannister, B. D. (1982). Effectiveness of performance feedback from behaviorally anchored rating scales. Journal of Applied Psychology, 67, 568-576.

Howe, E. S. & Silverstein, A. B. (1960). Comparison of two short-form derivatives of the Taylor Manifest Anxiety Scale. Psychological Reports, 6, 9-10.

Industry Day at the National Training Center. (1987). Orlando, FL: Project Manager for Training Devices, Naval Training Systems Center.

Isard, E. S. (1956). The relationship between item ambiguity and discriminating power in a forced-choice scale. Journal of Applied Psychology, 40, 266-268.

James, U. S., Pluger, W. D., & Duffy, P. (1983). Operational definitions of combat unit effectiveness and integrity (RN 83-46). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A136 757)

Johnson, J. H., Jones, M. B. & Kennedy, R. S. (1984). A video test for the prediction of tank commander performance. Proceedings of the 28th Annual meeting of the Human Factors Society, 502-503. Santa Monica, CA: Human Factors Society.

Johnston, W. A. (1966). Transfer of team skills as a function of type of training. Journal of Applied Psychology. 50, 102-108.

Kahan, J. P., Webb, N., Shavelson, R. J. & Stolzenberg, R. M. (1985). Individual characteristics and unit performance: A review of research and methods (MDA903-83-C-0047, TECH. REP. #R-3194-MIL). Santa Monica. CA: The Rand Corporation.

Kanarick, A. F. Alden, D. E. & Daniels, R. W. (1972). Decision making and team training in complex tactical training systems of the future. In V. Amico, J. Wooten and J. Regan (Eds). (1972). NTDC 25th Anniversary commemorative technical journal. Orlando, FL: Naval Training Device Center.

Kane, J. S. (1985). Rethinking the problem of appraising job performance, Unpublished paper.

Kane, J. S. & Lawler, E. E. (1979). Performance appraisal effectiveness: Its assessment and determinants. Research in organizational behavior (Volume I, pp. 425-478). New York: JAI Press, Inc.

Kavanagh, M. J., Hedge, J. W., Borman, W. C., Vance, R. J., Kraiger, K., Dickinson, T. & Banks, C. (1987). Issues and directions in job performance measurement. Symposium presented at the Annual Conference of the Society for Industrial and Organizational Psychology, Atlanta, GA.

Kay, B. R. (1959). The use of critical incidents in a forced-choice scale. Journal of Applied Psychology, 34, 269-270.

Kingstrom, P. O. & Bass, A. R. (1981). A critical analysis of studies comparing behaviorally anchored rating scales (BARS) and other rating formats. Personnel Psychology, 34, 263-289.

Knerr, C. M., Root, R. T. & Word, L. E., (1979). An application of tactical engagement simulation for unit proficiency measurement (Technical Paper 381). Alexandria, VA: U.S. Army Research Institute for the Behavioral & Social Sciences. (AD A074 410)

Kress, G. & McGuire, W. J. (1979). Implementation and evaluation of the tank crew training program for USAEUR units (ARI Research Note 79-40). Alexandria, VA: U.S. Army Research Institute for the Behavioral & Social Sciences. (AD A079 371)

Kribs, H., Thurmond, P., Mark, L. & Dewey, H. (1977). Computerized collective training for teams. (Interim Report DAHC 19-76-C-0042, ITR-77-A4). San Diego, CA: Sensors, Data, Decisions, Inc.

Kristiansen, D. M. & Witmer, B. G. (1981). Guidelines for conducting a training evaluation (TPE) (Research Product 81-17. Alexandria, VA: U.S. Army Research Institute for the Behavioral & Social Sciences. (AD A120 774)

Kubala, A. L. (1978). Problems in measuring team effectiveness. (Professional Paper No. 2-78, DAMC 19-75-C-0025). Alexandria, VA: Human Resources Research Organization (HumRRO).

Lahey, G. F. & Slough, D. A. (1982). Relationships between communication variables and scores in team training exercises. (Interim Report NPRDC-TR-82-35). San Diego, CA: Navy Personnel Research and Development Center.

Landy, F. J. & Farr, J. L. (1983). The measurement of work performance: Methods, theory, and applications. Orlando, FL: Academic Press, Inc.

Lane, N. E. (1986). Issues in performance measurement for military aviation with applications to air combat maneuvering. (Final Report DAAG29-81-D-0100, EOTR86-37). Orlando, FL: Essex Corporation.

Lane, N. E., Kennedy, R. S. & Jones, M. B. (1986). Overcoming unreliability in operational measures: The use of surrogate measure systems. Proceedings of the 1986 Annual Meeting of the Human Factors Society, 2, 1398-1402. Santa Monica, CA: Human Factors Society.

Latham, G. P. Fay, C. H. & Saari, L. M. (1979). The development of behavioral observation scales for appraising the performance of foremen. Personnel Psychology, 32, 299-311.

Lethan, G. P. & Wexley, K. N. (1977). Behavioral observation scales for performance appraisal purposes. Personnel Psychology, 30, 255-268.

Lintern, G., Nelson, B. E., Sheppard, D. J., Westra, D. P. & Kennedy, R. S. (1981). Visual technology research simulator (UTRS) human performance research: Phase III (NAUTRAEQUIPCEN 78-C-0060-11). Orlando, FL: Canyon Research Group.

MacDiarmid, R. (1987). Personal Communication.

Madden, J. M. & Bourden, R. D. (1964). Effects of variations in rating scale format on judgement. Journal of Applied Psychology, 48, 147-151.

Martinko, M. J. & Gardner, W. L. (1985). Beyond structured observation: Methodological issues and new directions. Academy of Management Review, 10(4) 676-695.

Meister, D. (1976). Team Functions. In Behavioral foundations of system development. New York: John Wiley & Sons.

Meliza, L. L. (1986). Defining roles in the development of a computer-aided ARTEP production system (CAPS) (Research Product 86-27). Alexandria, VA: U.S. Army Research Institute for the Behavioral & Social Sciences. (AD A178 445)

Miller, C. R. & Noble, J. L. (1983). Army training battle simulation system (ARTBASS) training development study. (TRASANATEA-60-82). Ft. Leavenworth, KS: U.S. Army Combined Arms Training Activity.

Mixon, T. R. (1982). An annotated bibliography of objective pilot performance measures, Part II (Technical Report No. NPS55-010PR). Monterey, CA: Naval Post Graduate School.

Morgan, B. B., Glickman, A. S., Woodard, E. A., Blaiwes, A. S. & Salas, E. (1986). Measurement of team behavior in a Navy training environment (Final Report NTSC TR-86-014). Norfolk, VA: Center for Applied Psychological Studies, Old Dominion University.

Muckler, F. A. (1977). Selecting performance measures: "Objective" versus "subjective" measurement. In L.T. Pope and D. Meister (Eds). Productivity enhancement: Personnel performance assessment in Navy systems. San Diego, CA: Navy Personnel Research and Development Center.

Murphy, K. R., Martin, C. & Garcia, M. (1982). Do behavioral observation scales measure observation? Journal of Applied Psychology, 67, 562-567.

Naval Training Equipment Center (1982). Surface ASW Training System (Device 141412): Functional Description (Contract No. 61139-80-D-0011). Orlando, FL: Author

Nebeker, D. M., Dockstader, S. L. & Vickers, R. R., Jr. (1975). A comparison of the effects of individual and team performance feedback upon subsequent performance (Final Report NPRDCTR 75-33). San Diego, CA: Navy Personnel Research and Development Center.

Nieva, V. F., Fleishman, E. A. & Rieck, A. (1978). Team dimensions: Their identity, their measurement and their relationships (Final Technical Report, DAHC19-78-C-001). Washington, DC: Response Analysis Corporation, Advanced Research Resources Organization.

Obermayer, R. W. & Vreuls, D. (1974). Combat-ready crew performance measurement (TECH. REP. #AFHRL-TR-74-108 (IV)). Northridge, CA: Manned Systems Sciences, Inc.

Obermayer, R. W. & Vreuls, D. (1974). Combat-ready crew performance measurement system: Phase II measurement system requirements (AFHRL-TR-74-108 (IV)). Northridge, CA: Manned Systems Sciences, Inc.

Obermayer, R. W. & Vreuls, D. (1974). Combat-ready crew performance measurement system: Phase IIIA measurement system requirements (AFHRL-TR-74-108(III)). Northridge, CA: Manned System Sciences, Inc.

Obermayer, R. E., Vreuls, D., Muckler, F. A., Conway, E. J. & Fitzgerald, J. A. (1974). Combat-ready crew performance measurement system: Final report (AFHRL-TR-74-108 (I)). Brooks Air Force Base, TX: Air Force Systems Command.

O'Brien, G. E., Crum, W. J., Healy, R. D., Harris, J. H. & Osborn, W. C. (1978). Trial implementations of the tank crewmans skills training program (TCST) (ARI TR 78-A29). Alexandria, VA: U.S. Army Research Institute for the Behavioral & Social Sciences. (AD A061 226)

Olmstead, J. A., Baranick, M. J. & Elder, B. L. (1978). Research on training for brigade command groups: Factors contributing to unit readiness (Technical Report 78-A18). Alexandria, VA: U.S. Army Research Institute for the Behavioral & Social Sciences. (AD A056 054)

Olmstead, J. A., Elder, B. L. & Forsyth, J. M. (1978). Organizational process and combat readiness: Feasibility of training organizational staff officers to assess command group performance (Technical Report 468). Alexandria, VA: U.S. Army Research Institute for the Behavioral & Social Sciences. (AD A138 881)

Pappa, S. C. (1986). Battlefield management system (BMS) concept overview. Proceedings Armor/Calvary Force in the 1980-90s Conference. Fort Knox, KY: U.S. Army Armor Center.

Patterson, D. G. (1922). The Scott Company graphic rating scale. Journal of Personnel Research, 1, 261-276.

Reaser, J. M. (1984). A methodology for and identification of command control group behaviors, Research Note 84-115).. Alexandria, VA: U.S. Army Research Institute for the behavioral & Social Sciences. (AD A145 690)

Report of Board of General Officers Appointed to Study Army Training Tests. (1 December 1959). Washington, DC: Headquarters, United States Army Corps.

Rosenblum, D. E. (1979). Combat effective training management study (CETRM). (Final Report October '78-July '79). Washington, DC: CETRM.

Roth, T. J., Hritz, R. J. & McGill, D. W. (1984). Model of team organization and behavior and team description method (Research Note 84-129). Alexandria, VA: U.S. Army Research Institute for the Behavioral & Social Sciences. (AD A147 540)

Ryan, T. G. & Yates, L. G. (1977). Report of exercise observations: Operational readiness training tests (OrTT) (Research Memorandum 77-7). Alexandria, VA: U.S. Army Research Institute for the Behavioral & Social Sciences. (AD A077 927)

Schrenk, L. P., Daniels, R. W. & Alden, D. G. (1969). Study of long-term skill retention (U) (Technical Report NAVTRADEVCEN 1822-1). Orlando, FL: Naval Training Device Center. CONFIDENTIAL.

Schwab, D. P., Heneman, H. G. III, & DeCotiis, T. A. (1975). Behaviorally anchored rating scales: A review of the literature. Personnel Psychology, 28, 549-562.

Semple, C. A., Cotton, J. C. & Sullivan, D. J. (1981). Aircrew training devices: Instructional support features. (AFHRL-TR-80-58). Wright-Patterson AFB, OH: Air Force Human Resources Laboratory.

Shaket, E., Saleh, J. & Freedy, A. (1981). Application of rule based computer models to the evaluation of combat training: A feasibility study (Research Note 81-13). Alexandria, VA: U.S. Army Research Institute for the Behavioral & Social Sciences. (AD A127 050)

Sharon, A. T., & Bartlett, C. J. (1969). Effect of instructional condition in producing leniency on two types of rating scales. Personnel Psychology, 22, 251-263.

Shiflett, J. E. (1986). Land battle test bed. Proceedings armor/calvary force in the 1980-90s conference. Fort Knox, KY: U.S. Army Armor Center.

Shiflett, S. C., Eisner, E. J., Price, S. J. & Schemmer, F. M. (1982). The definition and measurement of team functions. Final Report. Bethesda, MD: Advanced Research Resources Organization.

Smith, P. & Kendall, L. M. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scores. Journal of Applied Psychology, 47, 149-155.

Spencer, L. M., Jr. Klemp, G. O., Jr., Cullen, B. J. (1977). Work environment questionnaires and army unit effectiveness and satisfaction measures (Technical Report 491). Alexandria, VA: U.S. Army Research Institute for the Behavioral & Social Sciences. (AD A135 450)

Sulzen, R. H. (1980). The effects of repeated engagement simulation exercises on individual and collective performance (Technical Report 485). Alexandria, VA: U.S. Army Research Institute for the Behavioral & Social Sciences. (AD A100 974)

Swezey, R. W. (1978). Aspects of criterion-referenced measurement in performance evaluation. Human Factors, 20, 169-178.

Taylor, E. K., Schneider, D. E. & Clay, H. C. (1954). Short forced-choice ratings work. Personnel Psychology, 7, 245-252.

Thomas, G. S. (1984). Close air support mission: Development of a unitary measure of pilot performance (AFHRL-TR-84-39). Williams AFB, AZ: Air Force Human Resources Laboratory.

Thomas, G. S., Barber, H. F. & Kaplan, I. T. (1984). The impact of CATTS system characteristics on selected measures of battalion command group performance (Technical Report 609). Alexandria, VA: U.S. Army Research Institute for the Behavioral & Social Sciences. (AD A140 231)

Thorndike, R. L. (1949). Research problems and techniques. (Army Air Force Aviation Psychology Program Report No. 3.d. Washington, DC: Government Printing Office.

Thurmond, P. & Kribs, H. D. (1978). Computerized collective training for teams (ARI TR-78-A1). Alexandria, VA: U.S. Army Research Institute for the Behavioral & Social Sciences. (AD A050-890)

Turnage, J. J. & Muchinsky, P. M. (1982). Trans-situational variability in human performance within assessment centers. Organizational Behavior and Human Performance, 30, 174-200.

Turnage, J. J. & Lane, N. L. (1987). The use of surrogate techniques for the measurement of team performance. Proceedings of the 31st Annual Meeting of the Human Factors Society. Santa Monica, CA: Human Factors Society.

Turney, J. R. & Cohen, S. L. (1981). Defining the nature of team skills in Navy team training and performance (Final Report N0O0-14-80-C-0811, NR170-979). Columbia, MD: General Physics Corp.

Turney, J. R., Stanley, L., Cohen, S. L. & Greenberg, L. (1981). Targets for team skills training (Interim Report N00014-80-C-0811, NR170-919). Columbia, MD: General Physics Corp.

U.S. Army Training Board Improved ARTEP Briefing. (December 1984).

Vestewig, R. (1983). The SIMNET research program: Overview and performance measurement and system progress. Paper presented at the Military Testing Association conference, Washington, DC.

Vreuls, D. & Goldstein, I. (1976). In pursuit of the fateful few: A method for developing human performance measures for training control. In 9th NTEC/Industry Conference Proceedings (NAVTRAEQUIPCEN-IH-246). Orlando, FL: Naval Training Equipment Center.

Vreuls, D. & Obermayer, R. W. (1985). Human-system performance measurement in training simulators. Human Factors, 27(3), 241-250.

Vreuls, D., Obermayer, R. W., Woolridge, A. L., & Kelly, M. J. (1985). Team training and evaluation strategies; State-of-the-art (Technical Report, MDA903-76-C-0210). Alexandria, VA: Human Resources Research Organization (HumRRO).

Wagner, H., Hibbits, N. Rosenblatt, R. D., & Schultz, R. (1977). Team training and evaluation strategies; State-of-the-art (Technical Report, MDA903-76-C-0210). Alexandria, VA: Human Resources Research Organization (HumRRO).

Westra, D. P., Lintern, G., Sheppard, D. J., Thomely, K. T., Mauk, R., Wightman, D. C. & Chambers, W. (1986). Simulator design and instructional features for carrier landing. Transfer study (NAVTRASYSCEN 85-C-0044-2). Orlando, FL: Naval Training Systems Center.

Wherry, R. J. (1983). "Appendix: Wherry's Theory of Rating", In Landy, F. J. and Farr, J. L. The measurement of work performance: methods, theory and applications. Orlando, FL: Academic Press, Inc.

Zaval, A. (1965). Development of the forced-choice rating scale technique. Psychological Bulletin, 63, 117-124.

Zedeck, S. & Cascio, W. F. (1984). Psychological issues in personnel decisions. Annual Review of Psychology, 35, 461-518.